Abstract—Drug-drug interaction (DDI) detection is an important issue of pharmacovigilance. Currently, approaches proposed to detection DDIs are mainly focused on data sources such as spontaneous reporting systems, electronic health records, chemical/pharmacological databases, and biomedical literatures. However, those data sources are limited either by low reporting ratio, access issue, or long publication time span. In this work, we propose to explore online health communities, a timely, informative and publicly available data source, for DDI detection. We construct a weighted heterogeneous healthcare network that contains drugs, adverse drug reactions (ADRs), diseases, and users extracted from online health consumer-contributed contents, extract topological features, develop weighted path count to quantify the features, and use supervised learning techniques to detect DDI signals. The experiment results show that weighted heterogeneous healthcare network using leverage and lift are more effective in DDI detection than both unweighted homogeneous and heterogeneous network.

Keywords—drug-drug interactions; drug safety; pharmacovigilance; online health community; social media; heterogeneous network; weighted network; supervised learning

I. INTRODUCTION

Drug safety, also known as pharmacovigilance, is defined by the World Health Organization (WHO) as “the science and activities relating to the detection, assessment, understanding and prevention of adverse effects or any other possible drug-related problems” [1]. One important issue related to drug safety is how to detect signal of adverse drug reactions (ADRs). It has been long recognized that ADRs represent a significant worldwide health problem. In the United States, ADRs are considered to be a leading cause of death all over the country [2]. It is estimated that 82% of Americans are consuming at least one drug, 29% are consuming five or more drugs [3], and approximately 2 million patients in the United States are affected each year by ADRs [4]. Drug-drug interactions (DDIs), alterations of the effects of a drug due to the recent or simultaneous use of one or more other drugs, are another significant problem for drug safety, and they may account for up to 30% of unexpected ADRs [5]. In some extreme cases, adverse reactions caused by DDIs have led to death. For example, drug cerivastatin caused 31 cases of fatal rhabdomyolysis prior to June 2001, 12 of which involved the concomitant use of cerivastatin and gemfibrozil [5]. Consequently, DDI detection is also of great clinical importance and has been becoming an important research area in pharmacovigilance.

Currently, there are two main approaches to pharmacovigilance: pre-marketing and post-marketing surveillance. Before any new drugs are approved by Food and Drug Administration (FDA) for marketing, the pre-marketing review process is required. Some efforts have been made in clinical trials to detect DDIs [6, 7]. However, pre-marketing clinical trials are often constrained in terms of sample size, cohort biases, time spans, and financial limit to possibly identify all potential adverse reactions that may occur when the drug is used in clinical practice [8]. Furthermore, clinical trials primarily focus on ADR detection of single drugs and do not typically investigate DDIs [9]. Therefore, DDI detection depends heavily on post-marketing surveillance—the systematic detection and evaluation of medicines once they have been marketed.

The FDA Adverse Event Reporting System (FAERS), which contains over seven million reports of adverse events and reflects data from 1969 to the present, has become the primary data resource for the study and identification of new and unexpected postapproval ADRs in the United States [10]. Healthcare professionals and consumers voluntarily report to the system when they experience adverse events or medication errors. Although such systems could cover a large number of patients and drugs, it is well known that they are restricted by the problem of under-reporting, that is, the problem that not all occurrences of ADRs are reported to the spontaneous reporting systems [8]. It was reported that only 1 to 10 percent of all reportable adverse effects were reported to FAERS, and the majority of these reports came from drug companies [11]. Furthermore, it often takes FDA a long time to collect reports, investigate cases, and then release alerts. Therefore, it is usually impossible to detect adverse reactions timely by only relying on spontaneous reporting systems, thus endangering public health decisions. To sum up, it is urgent to find alternative data sources to supplement drug safety surveillance from spontaneous reporting systems.
Nowadays, the advancement of Internet not only breeds various online social networking sites such as Facebook, Twitter, etc., but also fosters online health communities (OHCs) such as MedHelp, WebMD, PatientsLikeMe, DailyStrength, and so forth. A recent survey conducted in September 2012 by Pew Internet & American Life Project showed that 72% of internet users said they went online for health information in the past year, 13% of which said they began their information seeking by visiting a site that specializes in health information, like WebMD [12]. We can easily imagine that uncountable health consumers and health professionals go to those OHCs frequently to either seek or offer healthcare information. For instance, since its introduction in 1994, MedHelp is the pioneer in OHCs. Today, MedHelp empowers over 12 million people each month to take control over their health and find answers to their medical questions [13]. Frequent visits on OHCs would inevitably generate mountainous health-related contents that might be more informative than some administrative databases. It is quite possible that many patients choose those online platforms to discuss adverse reactions they are experiencing. Increasingly, they share information, advice, and support via OHCs with professionals and even more often with fellow patients [14]. If we can take good advantage of these consumer-contributed contents, we may be able to reveal interesting and timely knowledge that may not be extracted from other data sources. Therefore, OHCs would serve as another data source to provide automated and augmented insight, discovery, and evidence-based health and wellness decision support. However, this data source still remains untapped by researchers in medical and healthcare domains [15].

In light of the popularity of social media in Web 2.0 and Health 2.0 era, it is beneficial to explore the potential of using OHC data for drug safety signal detection. Some of our previous studies have shown that OHC data can be used for pharmacovigilance [16-21]. Concretely, in [18], we proposed to perform association rule mining on consumer-contributed contents to identify associations between two drugs and an ADR. In [20], we utilized structural information of an unweighted heterogeneous network built from an OHC to detect drugs that were reported to interact with each other. However, in most real-world networks, relationships between different nodes are not entirely the same when links exist between them, which means binary representation is not sufficient to reveal the characteristics of links whereas link weights could carry richer information. Therefore, the goal of this study is to detect DDI signals from a heterogeneous healthcare network with considering link weights.

II. RELATED WORK

A. Drug-Drug Interaction Detection

1) Traditional Data Sources

Firstly, current post-marketing surveillance in United States primarily depends on FAERS, and many endeavors have been dedicated to detecting DDIs using reports from it, such as [22, 23]. Such systems have been proved to be a useful source to detect DDIs, but the effectiveness is constrained by high underreporting ratio. It is especially difficult to detect new and emerging signals because a large number of interesting cases cannot be timely collected due to underreporting ratio [24]. Also, long time span for regulatory agencies to release alerts is another issue.

Secondly, several studies proposed to mine DDIs from electronic health records (EHRs) such as [25, 26]. EHRs are primarily generated by health professionals in various healthcare organizations such as hospitals, clinics, etc. when, for example, a patient visits a doctor, therefore this kind of data is more timely and authoritative than spontaneous reports. However, EHR is often difficult to access because of privacy issues that it is usually available only to those research groups that have cooperation with hospitals, clinics or any other health organizations and communities. Also, the integration of heterogeneous electronic health databases is still a technical and policy challenge. Last but not the least, an analysis of an American Hospital Association’s recent survey found most American hospitals lag in implementing robust electronic health systems, and less than 6% of the hospitals meet all the requirements [27]. It means that EHRs also suffer the problem of not being robust and complete.

Thirdly, some studies also use chemical/pharmacologic databases to detect signals of DDIs. For example, DrugBank was used to identify novel DDIs [5]. Access issue is also one of the disadvantages of such data source because not all the chemical/pharmacological databases are free and publicly available to everyone. And they are more focused on the chemical aspect such as drug structure than textual aspect.

Last but not the least, natural language processing and text mining techniques applied to the biomedical literature can be of great benefit for pharmacovigilance, because it allows identification and extraction of drug-related information, and providing an interesting way to reduce the time spent by healthcare professionals and researchers who are trying to stay current by reviewing the literature [28]. Some studies took the advantage of such techniques for DDI signal detection, such as [29, 30]. Since published literature is often well-written with formal grammar, natural language processing and text mining techniques are often successfully used for knowledge extraction from such data source. However, the long cycle of publishing a paper in journals (months or even a year) could prevent timely drug safety signal detection.

2) Emerging Data Source – Consumer-Contributed Contents from Social Media

Although abovementioned traditional data sources have been widely utilized for DDI detection and abundant promising results have been shown, they still suffer from different kinds of limitations so that timely and effective drug safety signal detection will be hampered. It is urgent to find alternative data source to complement the traditional ones for better identifying DDI signals. As stated earlier, consumer-contributed contents from OHCs provide us with a great asset to mine knowledge and patterns. In the research area of DDI detection, such data source has also been used to achieve the goal although the number of such works is very limited. In 2013, White et al. demonstrated that Internet users are able to provide early clues about ADRs via their search logs [31]. In their study, they focused on Web search log data and paid particular attention to drugs paroxetine and pravastatin, whose interaction was reported to cause hyperglycemia. Then they used Reporting Odds Ratio to assess
the increased chance of a user searching for hyperglycemia-related terms given that they searched for both pravastatin and paroxetine. The experiment results showed that search logs of Web users can contribute to drug safety surveillance.

B. Link Prediction in Heterogeneous Information Network

An OHC itself is also a social network. In most of current research on network science, social and information networks are usually assumed to be homogeneous, where nodes are objects of the same type and links are relationships from the same relation type. However, most real-world networks are heterogeneous, where nodes and relations are of different types. For example, an OHC is a heterogeneous network. Not only does it include drugs and ADRs, but also contains other types of nodes such as diseases, treatments, users and so on. In the recent years, heterogeneous information network mining has been drawing increasing attention. Therefore, in addition to content analysis, structural analysis is another approach to discovering health-related knowledge.

Given the problem of DDI detection, we are actually predicting if there is an association between two drugs. Such problem can be formulated as link prediction that is an important task in network analysis. Link prediction is defined as predicting the emergence of links in a network based on certain current or historical network information [32]. Currently, most of the link prediction approaches are designed for homogeneous networks that only contain one type of objects such as authors in citation networks or social media users in friendship networks, and there are only a limited number of studies that have leveraged heterogeneous network for link prediction such as [33, 34].

To the best of our knowledge, there is no study that has leveraged heterogeneous healthcare network for DDI detection except our previous work [20]. In this paper, we combine the techniques and the data source, and propose to identify DDI signals by mining the structure of a weighted heterogeneous network that is extracted from OHCs.

III. METHODOLOGY

In this section, we introduce in detail the definition of heterogeneous healthcare network, the topological features extracted from such network, and the model for DDI detection task in such network setting.

A. Heterogeneous Healthcare Network

A heterogeneous network is defined as a graph $G = (\mathcal{N}, \mathcal{L})$ consisting of nodes joined by links, where $\mathcal{N} = \{n_1, n_2, \ldots, n_j\}$, $\mathcal{L} = \{l_1, l_2, \ldots, l_j\}$ and $l_i$ can be directional or non-directional. In the graph $G$, each node $n_i \in \mathcal{N}$ belongs to one particular type from $\mathcal{T}$, each link $l_i \in \mathcal{L}$ belongs to one particular relation from $\mathcal{R}$, and the number of the types of nodes $|\mathcal{T}| > 1$ or the number of types of relations $|\mathcal{R}| > 1$.

An OHC can be modeled as a heterogeneous network in which there are a set of node types, such as Drug, ADR, Disease, Treatment, Diagnostics, Users, etc. and a set of relation types, such as cause or is caused between Drug and ADR, treat or is treated between Treatment and Disease, use or is used between User and Drug, have or is had between User and Disease, etc.

B. Healthcare Network Model

A network model $M_G = (\mathcal{T}, \mathcal{R})$ is a compressed representation for a heterogeneous network $G = (\mathcal{N}, \mathcal{L})$, which is a directional or non-directional graph consisting of node types $\mathcal{T}$, with links as relations from $\mathcal{R}$. Fig. 1 succinctly presents a directional network model of a heterogeneous healthcare network. As we can see, the network includes four types of nodes, namely Drug, ADR, Disease, and User. For abbreviation, we use a capital letter to represent each node type, i.e. $R$ for Drug, $A$ for ADR, $D$ for Disease, and $U$ for User. The relations in this network contain cause or is caused between $R$ and $A$, treat or is treated between $R$ and $D$, show or is shown between $U$ and $A$, have or is had between $U$ and $D$, and take or is taken between $U$ and $R$.

A directional network model can be extracted from a heterogeneous network only when the relation between a pair of different types of node can be determined. For example, a bibliographic network can be represented by a directional network model. The relations among different types of node, such as paper, author, venue, and topic, can be explicitly and easily determined. Detailed examples of bibliographic heterogeneous network mining can be found in [33, 34]. However, not all heterogeneous networks contain explicit relations among different types of nodes, i.e. the semantic meaning of the relation could not be easily determined. Under such circumstances, the heterogeneous network could be represented as a non-directional network model and the relation between nodes can be the some kind of associations. For example, given a dataset of consumer-contributed, it is not an easy task to accurately determine the explicit relations between nodes without using sophisticated natural language processing (NLP) techniques or thorough human annotation. However, it is still challenging to use NLP tools to analyze social media data [35] and thorough human annotation would be very time consuming. In our work, we propose to analyze a non-directional heterogeneous healthcare network that contains 4 types of nodes (namely $R$, $A$, $D$, and $U$) joined together by their co-occurrence in an analysis unit. We can always expand the network by adding more types of nodes and relations in the future.

C. Topological Features in Heterogeneous Healthcare Network

Topological features are also called structural features, which are extracted connectivity properties for pairs of objects
in the networks [34]. Based on homogeneous network that only contains a specific type of nodes, there are a number of frequently used topological features. Most of the features are either path-based, such as graph distance, Katz [36] and propflow [37] or neighbor-based, such as common neighbors, Jaccard’s coefficient, Adamic/Adar, preferential attachment, and SimRank [36]. However, in a heterogeneous network, as a neighbor of one node could belong to different types and a path could also flow through different types of nodes, the commonly used features in homogeneous networks may no longer be applicable. For instance, in a heterogeneous healthcare network, two different drugs could be related by the path \( R - D - U - D - R \) because of the co-occurrence of each two adjacent nodes in analysis units, and the possible semantic meaning of such path could be explained as “a user has two different diseases which are treated by two different drugs respectively.” However, such information cannot be inferred from a homogeneous healthcare network that only consists of drugs. Therefore, some novel features that can reflect the characteristics of a heterogeneous network should be designed.

We define \( T_a T_d - Path - L \) as a topological feature of a heterogeneous network. A \( T_a T_d - Path - L \) is an abstract path defined between two types of nodes \( T_a \) and \( T_d \) with length \( L \). It is extracted from the network model \( M_G = (T, R) \), and is presented in the form of \( T_a \rightarrow T_1 \rightarrow \ldots \rightarrow T_{L-1} \rightarrow T_d \). When the specific types of relations and directions cannot be determined between nodes, \( T_a T_d - Path - L \) takes the form of \( T_a \rightarrow T_1 \rightarrow \ldots \rightarrow T_{L-1} \rightarrow T_d \) with links denoting associations between nodes. We extracted all the symmetric \( R(R_d) - Path \) with length 1 to length 4, and there are 16 such paths in total given 4 different types of nodes \( R, A, D, \) and \( U \), such as \( R-R, R-D-R, R-A-D-A-R, \) etc. (Table 1) The link existing between two nodes specifies the co-occurrence association between them.

**Table 1 Symmetric \( R(R_d) - Path \) to \( R(R_d) - Path \) in a Heterogeneous Healthcare Network**

<table>
<thead>
<tr>
<th>Path</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>R - R</td>
<td>1</td>
</tr>
<tr>
<td>R - A</td>
<td>2</td>
</tr>
<tr>
<td>R - D</td>
<td>2</td>
</tr>
<tr>
<td>R - U</td>
<td>2</td>
</tr>
<tr>
<td>R - A - R</td>
<td>3</td>
</tr>
<tr>
<td>R - D - R</td>
<td>3</td>
</tr>
<tr>
<td>R - U - R</td>
<td>3</td>
</tr>
<tr>
<td>R - A - A - R</td>
<td>4</td>
</tr>
<tr>
<td>R - A - D - A - R</td>
<td>4</td>
</tr>
<tr>
<td>R - A - U - A - R</td>
<td>4</td>
</tr>
<tr>
<td>R - D - A - D - R</td>
<td>4</td>
</tr>
<tr>
<td>R - D - D - D - R</td>
<td>4</td>
</tr>
<tr>
<td>R - U - D - D - R</td>
<td>4</td>
</tr>
<tr>
<td>R - U - A - U - R</td>
<td>4</td>
</tr>
<tr>
<td>R - U - D - U - R</td>
<td>4</td>
</tr>
<tr>
<td>R - U - U - U - R</td>
<td>4</td>
</tr>
</tbody>
</table>

**D. Weighted Heterogeneous Healthcare Network and Feature Quantification**

There are several ways of quantifying the topological features in a heterogeneous network. In [34], Sun et al. proposed to use such measures as path count, normalized path count, random walk, and symmetric random walk to quantify the features. There measures could also be applied into our non-directional heterogeneous network with some modifications. For example, we can use path count (PC) to measure the number of path instances between two nodes following a given \( T_a T_d - Path - L \).

Our previous study extracted topological features from an unweighted heterogeneous healthcare network and used path count to quantify them [20]. However, as stated earlier, in most real-world networks, relationship between different nodes are obviously not entirely the same when links exist between them, and link weights could carry more information. For example, given a heterogeneous healthcare network in Fig. 2, the number next to the link denotes the link frequency. If we don’t consider the weight, \( PC(R_1, R_2) \) under path \( R - A - R \) is 2, and \( PC(R_1, R_2) \) under path \( R - D - R \) is also 2. However, to some extent, \( path R - D - R \) is more interesting because the nodes under the path co-occurred more. Therefore, we assume that DDI detection based on a weighted heterogeneous network could achieve better performance. In this study, we propose to use three different metrics to weight the network: link frequency (LF), link leverage (LV), and link lift (LT).

Let \( l_{ab} \) be a link between nodes \( T_a \) and \( T_b \) and considering a thread of an OHC forum as an analysis unit, LF is the number of threads in which nodes \( T_a \) and \( T_b \) co-occur.

Leverage and lift are often used in association rule mining, one of the well-researched techniques in data mining. Association rule mining aims to extract interesting correlations, frequent patterns, associations or casual structures among sets of items in the transaction databases or other data repositories and is widely used in various areas such as telecommunication networks, market and risk management, inventory control, etc. [38] Mathematically, let \( I = \{l_1, l_2, \ldots, l_m\} \) be a set of items. Let \( X \), the task–relevant data, be a set of database transactions where each transaction \( T \) is a set of items such that \( T \subseteq I \). An association rule is an implication of the form \( A \Rightarrow B \), where \( A \subseteq I, B \subseteq I, \) and \( A \cap B = \emptyset \), where both \( A \) and \( B \) are a set of items, which is referred to as an itemset. Leverage and lift are often used to measure the interestingness and impressiveness of an association. In this study, we use these two to measure the importance of links of a heterogeneous network. Give a link \( l_{ab} \) between nodes \( T_a \) and \( T_b \), LV and LT are defined respectively as:

Fig. 2 An Example of A Heterogeneous Healthcare Network
\[ \begin{align*}
LV(t_{ab}) & = \text{support}(t_{ab}) - \text{support}(T_a) \times \text{support}(T_b), \\
LT(t_{ab}) & = \frac{\text{support}(t_{ab})}{\text{support}(T_a) \times \text{support}(T_b)},
\end{align*} \]

where

\[ \text{support}(t_{ab}) = \frac{LF(t_{ab})}{Z}, \quad \text{support}(T_a) = \frac{NF(T_a)}{Z}, \quad \text{support}(T_b) = \frac{NF(T_b)}{Z}, \]

where \(NF(T_a)\) and \(NF(T_b)\) denote node frequency of \(T_a\) and \(T_b\) respectively, and \(Z\) is the total number of threads in the dataset. For both LV and LT, the higher the value, the more importance the link will be.

After adding weight to the network, we propose to use Weighted Path Count (WPS) to quantify the extracted topological features. Given a \(T_s, T_d - Path - L\), the WPC is defined as:

\[ WPC(T_s, T_d) = \frac{1}{L} \sum_{P} \sum_{i=1}^{L} w(n_i, n_{i+1}) \]

where \(P\) denotes the path, \(L\) is the length of \(P\), \(n_i\) and \(n_{i+1}\) are two directly connected nodes following \(P\), and \(w(n_i, n_{i+1})\) is weight of the corresponding link connecting node \(n_i\) and \(n_{i+1}\).

Take the network in Fig. 2 as an example. If we use link frequency as the weight, (1) under path \(R - A - R\), \(WPC(R_1, R_2) = \frac{1}{2}(1 + 1 + 2 + 3) = 3.5\), and (2) under path \(R - D - R\), \(WPC(R_1, R_2) = \frac{1}{2}(3 + 4 + 5 + 6) = 9\). In this way, we can tell that for drug pair \(R_1\) and \(R_2\) path \(R - D - R\) has stronger association than \(R - A - R\).

E. Drug-drug Interaction Detection Model

In this study, we model DDI detection as a binary classification problem. Concretely, given a pair of drug nodes, we use a classification model to label them as either “1” (interaction) or “0” (no interaction) based on their quantified topological features extracted from the heterogeneous healthcare network. In this work, we are experimenting with three classifiers, namely Logistic Regression (LR), Naive Bayes (NB), and Support Vector Machine (SVM).

IV. EXPERIMENTS

In this section, we introduce in detail the dataset, network construction, gold standard, experiment setting, evaluation method, and experiment results.

A. Data Collection

In this study, MedHelp is used as the source of health-consumer-contributed contents. We focus on the drug section, which is one of the most important and popular components in MedHelp, and use the search engine I to search for all threads about a certain drug. To effectively detect DDI signals, the drugs should bear active discussion. Therefore, we targeted 20 drugs that have more than 500 threads for each of them, and collected all the original posts and following comments of those threads. The 20 drugs include Adenosine, Biaxin, Cialis, Concerta, Elidel, Epogen, Gadolinium, Geodon, Heparin, Lansoprazole, Lantus, Lunist, Luxvox, Prozac, Risperdal, Simvastatin, Tacrolimus, Vyvanse, Zocor, and Zyplar. The names of those drugs come from FDA’s website2, which includes an index of drugs that have been the subject of a Drug Safety Communication, Healthcare Professional Information sheet, Early Communication About an Ongoing Safety Review, or other important information. In total, there are 16,344 threads.

B. Network Construction

To construct the heterogeneous healthcare network, we need to extract different types of nodes and their relations. In this work, we focus on four types of nodes, namely \(R\), \(A\), \(D\), and \(U\), and external lexicons are used to extract them. For \(R\), besides the 20 drug names collected, we also add three other drugs (i.e., Quinidine, Ticlopidine, and Gemfibrozil) that could interact with some of the 20 drugs into our drug list. For more information about the three drugs and the interactions, please see [18]. For \(A\), we focus on 10 ADRs (i.e. Blurred Vision, Cancer, Depression, Diarrhea, Heart Disease, Hypertension, Kidney Disease, Skin Discoloration, Stroke, Suicide), and use Consumer Health Vocabulary (CHV) Wiki3 to build our ADR lexicon. More introduction of CHV can be found in [39]. CHV reflects the difference between consumers and professionals in expressing health concepts and helps to bridge this vocabulary gap. Therefore, high quality CHV is able to help with capturing more consumers’ expressions and better extracting ADR terms. Some studies are dedicated to expanding CHV by using social media data [40-42]. For \(D\), we search for diseases that are treated by each of the 20 drugs in SIDER database4 to construct our disease lexicon. SIDER contains information such as adverse drug reactions and diseases on marketed medicines, and the information is extracted from public documents and package inserts [43]. At last, there are 205 diseases in total, such as Bipolar Disorder, Hyperactivity Disorder, Hypercholesterolaemia, and so on. For \(U\), we extract all user names from each thread. The dataset is de-identified before conducting the experiment. For links, we treat our network as non-directional, and two nodes are linked together if they co-occur in the same thread.

C. Gold Standard

In the constructed healthcare network, the links between nodes are based on co-occurrence and their semantic meanings are implicit, so even if two different nodes are linked together, it does not mean that they will interact with each other. Therefore, an external database DrugBank is used to set up the gold standard. DrugBank database is a unique bioinformatics and cheminformatics resource that combines detailed drug data with comprehensive drug target information [44]. We search for all the 23 drugs to see if one drug is reported to interact with any other drugs using the Interax Interaction Search5 engine. If two drugs are reported to have interaction, we label the pair of drug nodes as “1”, and otherwise “0”. For example, Biaxin is

1 http://www.medhelp.org/search
2 http://www.fda.gov/Drugs/DrugSafety/PostmarketDrugSafetyInformationforPatientsandProviders/ucm111085.htm#B
3 http://consumerhealthvocab.chpc.utah.edu/CHVwiki/
4 http://sideeffects.embl.de/
5 http://www.drugbank.ca/interax/drug_lookup
reported by interact with Quinidine to cause Arrhythmias, Simvastatin is reported to interact with Gemfibrozil to cause Myopathy, etc.

**D. Experiment Setting**

In order to exclude the nodes and links that appear in the heterogeneous healthcare network rarely, we only retain the nodes and links with frequency larger than 15. After filtering, there are 511 nodes and 4378 links in our final network whose density is 0.034. For each pair of drug nodes, we use all the 16 symmetric $R_{i}R_{j} = \text{Path} - L$ as their topological features. Then we weigh the network by using LF, LV, and LT respectively and quantify the extracted features using WPC.

During the experiment, we found that our dataset is highly imbalanced, and the ratio of positive (drug pairs labeled as 1) and negative (drug pairs labeled as 0) instances is approximately 1:12. Therefore, we performed undersampling to build a new dataset with an equal sized set of positive and negative pairs. The experiment process is summarized as follows:

- **(1) Undersampling**: Given the positive drug pairs, randomly select an equal sized set of negative pairs to form a new dataset;
- **(2) Normalization**: normalize the new dataset to range [0, 1];
- **(3) Cross Validation**: perform 5-fold cross validation using multiple classifiers, i.e. LR (linear kernel), NB, and SVM (linear kernel);
- **(4) Loop**: repeat step (1) to (3) 5,000 times.

**E. Evaluation**

To evaluate the effectiveness of the proposed techniques, we set up two baselines for comparison:

- **(1) Comparison between heterogeneous and homogeneous networks**: We compare the performances between heterogeneous and homogeneous networks. Specifically, we constructed an unweighted homogeneous network that only contains one type of node – drug. We counted the number of path instances for each drug pair to quantify the homogeneous topological features with length no longer than 4, namely R–R, R–R–R, R–R–R–R, and R–R–R–R–R. Then we conducted the 4-step experiment as shown in section 4.4.
- **(2) Comparison between weighted and unweighted heterogeneous networks**: We compare the performances between weighted and unweighted heterogeneous networks. Specifically, we constructed an unweighted heterogeneous network, extracted all the 16 symmetric $R_{i}R_{j} = \text{Path} - L$ and used path count to quantify them. Then we conducted the 4-step experiment as shown in section 4.4.

We use F1 score and area under the ROC curve (AUC) to evaluate the proposed methods. Here the F1 and AUC scores are the average value of the 5,000 repeated experiments.

**F. Experiment Results and Discussion**

We first examine the performances of different classifiers in different networks. Fig. 3 and Fig. 4 illustrate F1 and AUC scores of different classifiers in different network settings respectively. Homo_Unweighted and Hete_Unweighted denote unweighted homogeneous and heterogeneous healthcare network respectively; Hete_LF, Hete_LV, and Hete_LT denote heterogeneous network weighted by LF, LV, and LT respectively. As we can see, in terms of both average F1 score and AUC score, the performance of different classifiers varies in different network settings. For example, for Homo-Unweighted, NB performed better than LR and SVM in terms of both F1 and AUC scores. However, for Hete_LF, SVM performed better than LR and NB in terms of both F1 and AUC scores. Also, for Hete_LV, SVM had the best performance in F1 score whereas LR outperformed NB and SVM in AUC score.

In general, we can observe that Hete_LV and Hete_LT outperformed two baselines and leverage-weighted network has the best performance. Then we conducted ANOVA analysis to see, for each classifier, if there is any significant difference between different network settings in terms of both F1 score and AUC score. Both Table 2 as well as Welch procedure ($p = .000$) and Brown-Forsythe procedure ($p = .000$) show that there is a statistically significant difference between different network settings in terms of F1 score if we use LR as classifier. We also tested NB and SVM, and the results show that a significant difference exists no matter which classifier we use. Furthermore, Games-Howell post-hoc tests demonstrate that for both F1 and AUC scores and under all three classifiers (1) Hete_LV is statistically significant higher than all other four network settings ($p = .000$ for all comparisons) except its AUC score is significant lower than all other four ($p = .000$) when NB is used, and (2) Hete_LT is statistically significant higher than Homo_Unweighted, Hete_Unweighted, and Hete_LF ($p = .000$ for all comparisons) except that its AUC score is significant lower than Homo-Unweighted ($p = .000$), Hete_Unweighted ($p = .000$), and Hete_LF ($p = .000$) when we use NB ($p = .000$). Furthermore, ANOVA analysis under network setting Hete_LV showed that (1) for F1 score, NB is statistically significant higher than LR ($p = .000$) and SVM ($p = .000$), but there is no significant difference between LR and SVM ($p = .075$); (2) for AUC score, both LR and SVM are significant higher than NB ($p = .000$), but there is no significant difference between LR and SVM ($p = .279$).

<table>
<thead>
<tr>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Groups</td>
<td>1401.278</td>
<td>4</td>
<td>350.319</td>
<td>9223.92</td>
</tr>
<tr>
<td>Within Groups</td>
<td>949.296</td>
<td>24995</td>
<td>.038</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>2350.574</td>
<td>24999</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The results demonstrate that (1) the performance of different classifiers varies under different evaluation scenarios, and (2) leverage- and lift-weighted heterogeneous healthcare networks are generally more effective in DDI detection than both unweighted homogeneous and heterogeneous networks, as}

---

*Note: The table and figures are not included in the text representation.*
well as frequency-weighted heterogeneous network. Link frequency is proportional to its support value \( \text{support}(l_{ab}) = \frac{\text{support}(A \cup B)}{A \cup B} \), which is not enough to represent the information that the link carries. For example, considering a link \( R \rightarrow A \), its support value could be very small, but it does not mean that this link is trivial, because this ADR could be one of the rare ADRs that would be caused by the drug. One limitation of support is that it would work well when the ADR of the drug appear frequently in the dataset. However, health consumers discuss diverse aspects of drugs in online forum, such as drug dosage, drug prescription, concomitant use of different drugs, and so forth. It is very likely that threads that are related to the specific ADR are only a small portion of the total threads, especially for those rare ADRs. Leverage and lift could be used to address this problem because they incorporate the support of the ADR in the dataset. Both leverage and lift measure the strength of a link not only by looking at its support but also the correlation between the two nodes. Leverage measures the difference between the proportion of threads containing both nodes above those expected if the two nodes were independent of each other whereas lift calculate the ratio of these two. Therefore, both leverage- and lift-weighted heterogeneous network perform better than frequency-weighted one.

![Fig. 3 F1 Scores of Different Classifiers in Different Network Settings](image)

![Fig. 4 AUC Scores of Different Classifiers in Different Network Settings](image)

V. CONCLUSION

The development of Health 2.0 technologies leads the booming of OHCs such as MedHelp, WebMD and so on. Such platforms are not only empowering individuals to play a substantial role in their own health, but also generating informative data that can be used to provide automated insights and discovery. DDIs are a serious drug safety problem for health consumers and how to detect DDI signals effectively and efficiently has been of great medical significance. Currently, methods proposed to detect DDIs are mainly based on such data sources as spontaneous reporting systems, electronic medical records, chemical/pharmacological databases, and pharmacovigilance literatures. However, those sources are limited either by low reporting ratio, access issue, or long publication cycle. In this study, we proposed to harness OHC data for DDI detection. We used MedHelp as our source to collect consumer-contributed contents based on which a weighted heterogeneous network was constructed. Then we extracted topological features from the network, quantified them with different weighting methods, and used various classifiers for DDI signal detection. The experiment results showed that the proposed weighted heterogeneous network substantially outperformed the unweighted homogeneous and heterogeneous counterparts. In the future, this work can be extended in several directions: (1) more types of nodes could be added into the heterogeneous network such as diagnoses, treatments, etc.; (2) various co-occurrence levels can be considered, such as post and sentence; (3) asymmetric paths could also be considered as topological features; (4) other than using undersampling to build a balanced dataset, more machine learning techniques could be considered such as oversampling, cost-sensitive learning, ensemble learning, etc.; and (5) based on the prediction, we can tell which two drugs will interact with each other, but one important thing missing is what adverse reaction these two drugs could cause due to interaction. In the future, we propose to extend our research along this direction using triad prediction. A triad is a group of three nodes and it is one of the most basic units for studying group phenomena in social networks [45]. Given a heterogeneous healthcare network, two drugs and an ADR can be regarded as a triad and triadic formation can be studied using prediction approaches based on a set of features of the triad, such as number of links, triad type, link weight distribution, node role, and so on.

REFERENCES


[35] K. Denecke, "Extracting Medical Concepts from Medical Social Media with Clinical NLP Tools: A Qualitative Study."