Exploiting OHC Data with Tensor Decomposition for Off-label Drug Use Detection

Mengnan Zhao
College of Computing and Informatics
Drexel University
Philadelphia, PA, US
mz438@drexel.edu

Christopher C. Yang
College of Computing and Informatics
Drexel University
Philadelphia, PA, US
chris.yang@drexel.edu

***Abstract*: Off-label drug use is an important healthcare topic as it is quite common and sometimes inevitable in medical practice. Though gaining information about off-label drug uses could benefit a lot of healthcare stakeholders such as patients, physicians, and pharmaceutical companies, there is no such data repository of such information available. There is a desire for a systematic approach to detect off-label drug uses. Other than using data sources such as EHR and clinical notes that are provided by healthcare providers, we exploited social media data especially online health community (OHC) data to detect the off-label drug uses, with consideration of the increasing social media users and the large volume of valuable and timely user-generated contents. We adopted tensor decomposition technique, CP decomposition in this work, to deal with the sparsity and missing data problem in social media data. On the basis of tensor decomposition results, we used two approaches to identify off-label drug use candidates: (1) one is via ranking the CP decomposition resulting components, (2) the other one is applying a heterogeneous network mining method, proposed in our previous work [9], on the reconstructed dataset by CP decomposition. The first approach identified a number of significant off-label use candidates, for which we were able to conduct case studies and found medical explanations for 7 out of 12 identified off-label use candidates. The second approach achieved better performance than the previous method [9] by improving the F1-score by 3%. It demonstrated the effectiveness of performing tensor decomposition on social media data for detecting off-label drug use.**

Keywords: off-label drug use, online health community, heterogeneous network, tensor decomposition

I. INTRODUCTION

Off-label drug use refers to the use of marketed medications on the indications that are not on their FDA-approved labeling information. Although Food and Drug Administration (FDA) is responsible for approving new medications to the market, they do not manage or supervise whether the drugs are used as directed in the administration.

Off-label drug use is very common in medical practice, including both over-the-counter drugs (OTCs) and prescription drugs. For instance, up to one-fifth medications are prescribed off-label [1]. Off-label uses are more frequent for some special subpopulations, such as younger children, gravida, and the elderly, because they are often excluded for drug test or drug screening, thus leading less drugs approved for them exclusively [2]. Off-label uses are also more frequent for some specific medications, for example, the antipsychotic drugs approved for one psychiatric disorder are sometimes prescribed for the other psychiatric disorders [2].

Off-label drug uses are sometimes beneficial for patients, whereas a majority of them lack sufficient scientific research or test. Healthcare providers usually prescribe off-label drugs for the benefit of patients based on their experience, and some drugs are even used in off-label ways more often than their approved indications and deliver effective treatment, but still a majority of off-label uses have no or less scientific evidence [1], which could cause serious outcomes, adverse effects, or medication errors.

As off-label drug uses are almost inevitable in medical practice, it puts forward the need to detect such uses in a systematic way, which benefits most of the healthcare-related stakeholders. Specifically, healthcare providers and patients can gain information about the observation of off-label drug use in practice, pharmaceutical companies can rely on such information to create postmarketing surveillance reports of medications, and biomedical researchers may find new practicable drug uses or potential risks from the existing off-label use practice. Despite of the need to detect off-label drug uses, there have not been effective ways to acquire and maintain such information. Survey has been a popular method to identify off-label drug uses, but limited by many conditions, such as the number of physicians or patients that respond, the size and quality of collected data, and the extensive time and cost that involved.

Recently, the wide use of electronic health record (EHR) systems has enable researchers to gain information of off-label drug use in a scalable manner [3], whereas such approaches could be limited by the accessibility of hospital data, the complexity of EHR data, and the interoperability of different EHR systems. Meanwhile, the popularity of social media, in particular online health communities (OHCs), suggests a promising way to obtain related information from health consumers directly. There is an increasing number of healthcare consumers exchanging information about the medications and treatments through OHC, and hence generating a large volume of accessible and valuable information. OHC data have been successfully applied to many healthcare issues such as the detection of drug-drug interactions [4] and adverse drug reactions [5-7]. Moreover, some previous work has also shown the feasibility to detect off-label drug uses from OHC data [8-9].

Although OHC data contain large amounts of health-related information, it also poses new challenges. OHC data suffers in

2575-2634/18/$31.00 ©2018 IEEE
DOI 10.1109/ICHI.2018.00010

problems such as misspellings and omissions. Sparsity is another well-known problem for OHC or social media data. Missing data problem is common in many areas such as social network analysis, image processing, and data mining. Our previous work also suffered this problem when creating multidimensional datasets from the OHC data. For example, in the work of detecting repositioning drugs [10], we created a three-dimensional dataset that measures the co-mention of disease, drug, and adverse drug reaction (ADR), where a large number of the cells were filled with 0, due to the noisy and sparse social media data. If we can well handle the missing values, we may be able to achieve better results. Tensor decomposition is a popular tool used to handle missing data because it is effective in capturing dependencies in high-dimensional datasets [11]. In this work, we propose a tensor-based method to exploit social media data for off-label drug use detection.

II. RELATED WORK

Survey has been the most popular method to study off-label drug uses and related topics, such as the number or ratio of off-label drug uses in practice and the associations between off-label drug use and potential adverse drug effects [12-15]. For instance, a prospective study about the off-label drug use for children in Europe has been conducted and the results showed that 39% of the prescriptions involved off-label drugs [16].

Recently, with the digital availability of medical documents, some studies have exploited electronic data sources such as EHR and clinical notes for the detection of off-label drug uses [17-18]. Inferring novel drug-disease relationships is the most critical step in detecting off-label drug use and has been studied in a lot of researches. Some researches employed text mining and natural language processing (NLP) based methods such as lexicon-based approach [19] and word embedding algorithm [8], whereas more researches utilized network-based methods such as random walk [20], propagation flow [21-23], and meta-path-based algorithm [9].

While most off-label drug use researches used data sources from healthcare providers, only a few researches have utilized the data contributed by health consumers. It has been claimed that 80% of adults in the US and 66% of adults in Europe seek online health advice [24]. There is an increasing number of Internet users engaging in online health communities to discuss their health conditions, share and confirm information [25], seek emotional support [26], and get social bonds [27]. OHC data thus contains a large volume of health-related information. However, sparsity and missing data problems have been the main challenge of exploiting social media or OHC data. Tensor decomposition is a popular technique for handling sparsity and missing data issues, which is quite appropriate to address multidimensional social media data and has been successfully applied to many healthcare topics. For instance, tensor decomposition has been utilized to fill missing data in medical questionnaires [28], to localize and extract artifacts from EEG data [29-30], and to identify adverse drug reactions for pharmacovigilance [31].

III. OFF-LABEL DRUG USE DETECTION WITH TENSOR DECOMPOSITION

In this section, we present the proposed approach of exploiting OHC data with tensor decomposition to detect off-label drug uses. First, we collected data from the OHC website and preprocessed the raw text data, with the goal to locate the involved entities (i.e. drug, disease, user) and compute the co-mentions. Secondly, we created a three-dimensional dataset and conducted tensor decomposition on the three-dimensional dataset. Lastly, we used two approaches to identify off-label drug uses: (1) we ranked the obtained tensors to identify candidates for off-label drug use; (2) we reconstructed the multidimensional dataset from our decomposition, constructed a heterogeneous healthcare network based on the reconstructed multidimensional dataset, and then employed the meta-path-based approach [9] proposed in our previous work to identify off-label drug uses. Fig. 1 presents the framework of the proposed approach.

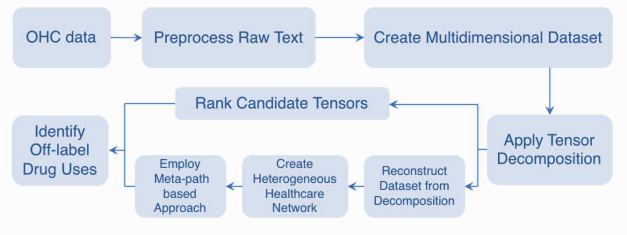

Fig. 1 Flow of exploiting OHC data with tensor decomposition for off-label drug use

A. Dataset and preprocessing

The era of Web 2.0 fosters not only a number of social media websites such as Facebook and Twitter, but also many online health communities such as MedHelp, PatientsLikeMe, and DailyStrength. Increasing people discuss their health issues or seek emotional support in OHCs. MedHelp (www.medhelp.org), as a pioneer of OHC website, owns 176 disease communities and attracts more than 12 million visitors each month to post on their site, which generates a large volume of health-related contents. When most previous studies about off-label drug use detection utilized EHR, clinical notes and the other data source provided by healthcare professionals, we have explored the possibility of taking advantage of those consumer-contributed contents that are valuable, interesting, and more timely.

We collected data from MedHelp with an automatic web crawler and retrieved posts and comments in the most popular 50 disease communities, obtaining about 71,000 threads (71,000 posts + 320,000 comments). For each of the collected thread, it contained the original post, the following comments, and their writers' usernames.

Since the goal was to create a three-dimensional dataset of disease, drug, and user, we first extracted these entities from the collected text corpus. For the extraction of diseases and drugs, we referred to several databases (i.e. UMLS and PharmGKB) to identify them. Unified Medical Language System (UMLS) is a medical vocabulary and terminology system maintained by US National Library of Medicine;

Pharmacogenomics Knowledgebase (PharmGKB) is a publicly available data resource including knowledge of human genetic variation on drug responses, gene-drug associations, and drug labels. In addition, considering health consumers may mention diseases and drugs with non-professional vocabulary, we also resorted to Consumer Health Vocabulary (CHV) Wiki to identify them. CHV Wiki is a data repository that connects the professional expressions of medical terms with the everyday expressions. For example, "Parkinson's disease" is the professional term for this disease, CHV extends it to "PD" "Parkinson" "Parkinsons" and several other common expressions from health consumers. For the extraction of usernames, we de-identified them with random IDs for the privacy protection.

B. Multidimensional dataset construction

After identifying the necessary entities (i.e. drug, disease, user) in the text, we constructed a three-dimensional dataset that embodied the co-mentions between drugs, diseases, and users. For each cell, it computed the number of times that the corresponding drug and disease were both mentioned by the corresponding user in a comment or post.

In the current stage, we focused on the 50 diseases in the MedHelp communities and the 1,297 drugs indicated for those diseases by PharmGKB, and then detected about 193,000 users from the text. In order to manage the size of the dataset and prepare an appropriate sparsity for tensor decomposition, we removed the users that have no connection with any drug or disease and the drugs that have no connection with any user. After the removal, the dataset has 313 drugs and 2465 users, with sparsity equal to 0.0004. In addition, we also removed users who published less than 10 comments/posts so as to ensure we got adequate information of each user. 409 users were retained in the dataset. We created a three-dimensional dataset with the size of (50, 313, 409) and the sparsity of 0.022.

C. Tensor decomposition

Tensor decomposition is an important technique for handling multi-dimensional data, which is capable of dealing with the problem of missing data and modeling those dimensions simultaneously without collapsing the dataset into matrices [32]. Tensor decomposition was utilized to explore our three-dimensional dataset here, allowing us to fill out the missing value and to identify off-label drug use candidates via analyzing the connection between three factors: drug, disease, and user.

A tensor is a multidimensional array. The *order* of a tensor is the number of dimensions. For instance, a first-order tensor is a vector, a second-order tensor is a matrix, and a *N*th-order tensor is made up of *N* vector spaces ($\boldsymbol{\chi} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$). The tensor of order three or higher is named higher-order tensor. If a *N*th-order tensor can be written as the outer product of *N* vectors, i.e. $\boldsymbol{\chi} = \boldsymbol{a}^{(1)} \circ \boldsymbol{a}^{(2)} \circ \cdots \circ \boldsymbol{a}^{(N)}$ ($\circ$ denotes the outer product, $x_{i_1 i_2 \cdots i_N} = a_{i_1}^{(1)} a_{i_2}^{(2)} \cdots a_{i_N}^{(N)}$), the tensor is a *rank-one* tensor.

Tensor decomposition or factorization is a natural extension of matrix factorization that can capture underlying patterns in multi-dimensional datasets and have applied successfully in many disciplines [33]. One of the most popular tensor decompositions is CANDECOMP/PARAFAC (CP) decomposition, proposed by Carroll and Chang [34] and Harshman [35]. The reason we utilized CP decomposition here is because it has been successfully used in medical studies to fill missing data [28] and the structure of results, *rank-one* tensor, is appropriate for capturing and representing medical concepts and easy to interpret [36].

CP decomposition expresses a tensor as the linear combination of a finite number of *rank-one* tensors. The smallest number of *rank-one* tensors is defined as the rank of the tensor. Given a third-order tensor $\boldsymbol{\chi} \in \mathbb{R}^{M \times N \times J}$ and assume its rank as *K*, then its *K* component CP decomposition is written as:

$$\boldsymbol{\chi} \approx \sum_{k=1}^{K} \lambda_k \boldsymbol{a}_k \circ \boldsymbol{b}_k \circ \boldsymbol{c}_k$$

in which, *K* is a positive integer, λ is the weight of each component, and $\boldsymbol{a}_k \in \mathbb{R}^M$, $\boldsymbol{b}_k \in \mathbb{R}^N$, and $\boldsymbol{c}_k \in \mathbb{R}^J$. Usually, $\| \boldsymbol{a}_k \| = \| \boldsymbol{b}_k \| = \| \boldsymbol{c}_k \| = 1$ and λ_k contains the scalar weight. For each element in $\boldsymbol{\chi}$,

$$x_{mnj} \approx \sum_{k=1}^{K} \lambda_k a_{mk} b_{nk} c_{jk} \quad for\ m = 1, \ldots, M; n = 1, \ldots, N;\ j = 1, \ldots, J$$

Due to missing data and noise, the true $\boldsymbol{\chi}$ is not observable and equality cannot be expected, therefore, the CP decomposition should achieve approximation by minimizing the following cost function:

$$\min J = \frac{1}{2} \sum_{m=1}^{M} \sum_{n=1}^{N} \sum_{j=1}^{J} (x_{mnj} - \sum_{k=1}^{K} \lambda_k a_{mk} b_{nk} c_{jk})^2$$

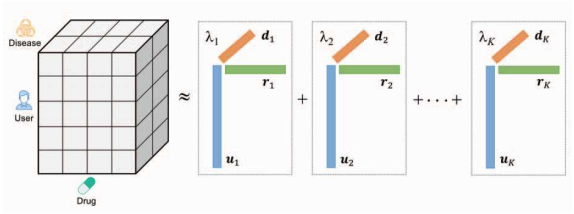

Fig. 2 Generating off-label candidates using CP decomposition

In this work, we transform our dataset into a third-order tensor $\boldsymbol{\chi} \in \mathbb{R}^{D \times R \times U}$, where *D*, *R*, *U* is the number of diseases, drugs, and users. Each tensor element denotes the number of times drug *r* is co-mentioned with disease *d* in all the posts and comments of user *u*. This third-order tensor can be interpreted from three views: (1) user view: a matrix of the user's diseases and drugs, (2) drug view: a matrix of the users and the diseases associated with this drug, (3) Disease view: a matrix of the users and the drugs associated with this disease. In addition, using tensor to represent the interactions between drug, disease, and user allows to only store non-zero elements for efficient memory storage. Fig. 2 conceptually illustrated the process of a CP decomposition on our third-order co-mention tensor, in which, $\boldsymbol{d}_k$, $\boldsymbol{r}_k$, and $\boldsymbol{u}_k$ represent *disease vector*, *drug vector*, and *user vector* respectively.

D. Identification of off-label drug use

1) Candidate identification via ranking resulting components

Applying CP tensor decomposition on a third-order tensor generates three factor matrices. In our work, they are matrices of *disease factor* (***D***), *drug factor* (***R***), and *user factor* (***U***), with size of $D \times K$, $R \times K$, $U \times K$ respectively. The elements in the matrices denotes the probability of seeing that element in the corresponding component. For instance, in disease factor matrix $\boldsymbol{D}$, $d_{ik}(i = 1, \dots, D, k = 1, \dots, K)$ denotes the probability of seeing disease d_i in resulting component k. From another viewpoint, CP tensor decomposition derives K resulting components, where each component contains the scalar weight λ, *disease vector* $\boldsymbol{d}$, *drug vector* $\boldsymbol{r}$, and *user vector* $\boldsymbol{u}$. Moreover, since λ represents the significance and the ability to capture the tensor data of that component, it can be used to automatically rank the components in order of significance. In addition, we set a hard thresholding constraint to identify the significant entities (i.e. disease, drug, and user) in each vector. In sum, three steps were taken here to identify off-label drug use candidates from results of CP decomposition: (1) identify significant components by ranking them based on λ; (2) identify significant entities in vector $\boldsymbol{d}$, $\boldsymbol{r}$, and $\boldsymbol{u}$ by employing a hard-thresholding constraint; (3) identify off-label drug use candidates by removing the on-label drugs.

2) Candidate identification via meta-path-based methods

Tensor decomposition is a good technique to be used for handling missing data problem, as it is able to capture the dependencies in multidimensional dataset and to learn the latent and collaborative relationship structure from different perspectives (i.e. dimensions, factors) [37]. The original tensor with missing data issue can be filled and reconstructed from the decomposition results. Here we reconstructed our three-dimensional tensor from the CP decomposition results.

Based on the recovered full tensor $\boldsymbol{\chi}'$ that contains co-mention counts between disease, drug, and user, we could measure the co-occurrence frequencies and create a heterogeneous healthcare network, and then applied the meta-path-based method proposed in our previous work [9] to identify off-label drug uses. Briefly, there are three main steps:

(1) Compute the recovered co-mention frequency between disease and drug, disease and user, and drug and user from the recovered tensor $\boldsymbol{\chi}'$. For example, given disease $d_i$and drug r_j, the co-mention frequency, $c(d_i, r_j)$, were computed by the following equation:

$$c(d_i, r_j) = \sum_{u=1}^{U} x'_{iju}$$

Similarly,

$$c(r_i, u_j) = \sum_{d=1}^{D} x'_{dij} \qquad c(d_i, u_j) = \sum_{r=1}^{R} x'_{irj}$$

(2) Measure the strength of co-mention frequency with *lift*, a popular indicator from association rule mining based on probability and reflecting the division of the actual probability and theoretical probability.

(3) Construct a heterogeneous health network, determine meta-paths and compute the weights of meta-paths based on *lift*. The network contained three types of nodes (i.e. disease, drug, and user) and three types of edges (i.e. disease-drug, drug-user, and disease-user), as shown in Fig. 3. Then five meta paths were derived from that network.

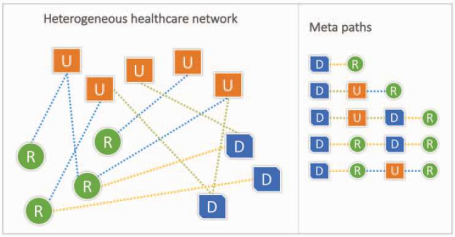

Fig. 3 The heterogeneous healthcare network including drug, disease, and user

(4) Train a classifier based on Random Forest algorithm to distinguish known drug-disease pairs with those unknown by using the extracted weights as input features and identify the off-label drug uses from the classification results.

IV. EXPERIMENT & RESULTS

In this work, we exploited social media data to identify off-label drug uses, moreover, we adopted tensor decomposition technique in order to resolve the sparsity and missing data problems. As described in section III, we first retrieved the consumer-contributed data from MedHelp and collected about 71,000 posts and 320,000 comments, from which, we detected 50 diseases, 1,297 drugs and about 193,000 users. Secondly, we filtered out a number of drugs and users with consideration of data size, sparsity, and computation cost, which generated a three-dimensional dataset of 50 diseases, 313 drugs, and 409 users. Thirdly, we utilized the CP decomposition algorithm by using TensorLy package for Python [38]. We identified off-label drug use candidates from CP decomposition results.

A. Results of ranking CP decomposition components and case study

Applying CP decomposition generated K resulting components, where K is a parameter that should be set in advance. Based on the tensor size and computation time, we set K = 100 and ranked the components in descending order according to the value of λ. Then we identified the off-label candidates from each component. Fig. 4 illustrated an example of identifying off-label candidates from a component, in which, six drugs and one disease, myalgia, occurred together and the percentage of users were calculated by counting the non-zero elements in the *user vector*. Among the six drugs, methotrexate is indicated for myalgia and aspirin is a common pain reliever, so the other four drugs are potential to be the off-label medications for myalgia.

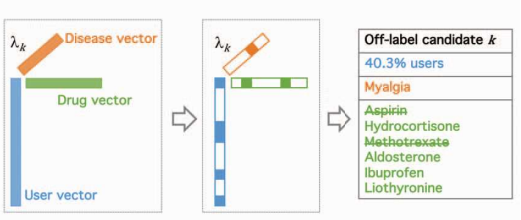

Fig. 4 An example of identifying the *k*th off-label candidate identified via ranking resulting components

The above illustration is one of the resulting components in tensor decomposition. We have conducted a case study on the most significant off-label candidate, which is the resulting component with the highest weight. The highest weighted (largest λ) candidate of off-label drug use contains two diseases (hypertension and Parkinson's disease (PD)), six off-label drugs(ibuprofen, methadone, rivastigmine, lactulose, boceprevir, and citalopram) after removing three indicated medications (acebutolor, benazepril, and methadone), and 53.5% users.

The association between hypertension and PD is easy to understand as they both have higher rates to present on patients of older age, while the underlying pathology remains unknown [39].

Ibuprofen is a nonsteroidal anti-inflammatory drug used for treating pain, fever and inflammation. There has been a hypothesis that anti-inflammatory drugs like ibuprofen may contribute to the pathogenesis of PD and thus reduce PD incidence [40], therefore, there is the possibility that ibuprofen is used for PD in off-label way. While ibuprofen may induce the blood pressure to rise, so it is risky for hypertension patients to use ibuprofen.

Rivastigmine is originally indicated for Alzheimer but sometimes used on PD patients to treat the cognitive impairment [41]. However, we cannot find any evidence between rivastigmine and hypertension.

Lactulose is a synthetic sugar indicated for constipation and hepatic encephalopathy. Lactulose may induce high blood pressure for hypertension patients [42]. For PD, it does not provide effective treatment, but may help to improve mental status [43].

Citalopram is an antidepressant, which is thought to be effective for PD by improving response inhibition in some studies [44], but it is also found to be risky due to the possibility of exacerbating abnormal movements [45]. As depression occurs at a higher rate in patients with hypertension than the normal people and depression induce poor hypertension control [46], citalopram might be prescribed for patients with hypertension especially the elderly patients.

Methadone is an opioid medication used as pain reliever or maintenance therapy of drug addiction detoxification. Boceprevir is an antiviral medication used to treat chronic Hepatitis C. Though we did not find enough medical evidence to support the relationship between methadone/boceprevir and hypertension/PD, there was still the possibility that they occur together in clinical practice a lot.

B. Results of applying meta-path-based method on the reconstructed tensor

In our previous study [9], we have developed an automated method to identify off-label drug use from OHC data based on heterogeneous network mining. The basic idea is constructing a heterogeneous healthcare network based on the calculation of co-occurrence frequency, deriving meta-path features from the network, and then utilizing the features to train classifiers for the identification of off-label drug uses.

In the previous research [9], we calculated the co-occurrence frequency directly from the count matrices without dealing with the missing data problem. In this work, after obtaining the original count matrices, we applied tensor decomposition (TD) to learn the latent structure and dependencies of the matrices and then reconstructed the count matrices from tensor decomposition results. The following approaches were the same with what we have performed in the previous study [9]. We evaluated the proposed tensor decomposition approach (*TD-PCL*) with the best approach in the previous paper, that is, Path-Count-Lift (*PCL*).

The dataset for classification contained 491 positive drug-disease pairs and 8,000 randomly-generated negative pairs, then we performed undersampling to avoid the imbalance problem and 10-fold cross validation to train and test the Random Forest based classifier. The classification performance was evaluated by Precision, Recall and F1-score, as shown in Table 1.

Table 1 Classification evaluation results on TD-PCL and PCL

Feature	Dataset	Random Forest		
		Precision	Recall	F1 score
PCL	Training	0.833	0.942	0.884
	Test	0.724	0.869	0.790
TD-PCL	Training	0.852	0.939	0.893
	Test	0.748	0.895	0.815

The results showed that using the heterogeneous network-based method to classify disease-drug associations were effective and achieved a promising performance, with or without adopting tensor decomposition. While *TD-PCL* outperformed *PCL* in terms of the indicators, for instance, F1-score, the harmonic average of Precision and Recall, was improved 3%. The results indicated the advantage of performing tensor decomposition before constructing the heterogeneous healthcare network. *TD-PCL* achieved Recall of 0.895, denoting that among all the known disease-drug usages 89.5% were classified correctly, and achieved Precision of 0.748, representing that 74.8% of the predicted positive drug-disease relationships were indeed known usages and the other 25.2% were falsely classified as positive. The hypothesis here is that if the network features extracted from OHC data could enable us to recognize the known disease-drug pairs, then the other pairs that show similar features with the known pairs, which are the false-positive pairs, have a high potential to be the off-label practices [9]. We identified the potential off-label drug-disease pairs from the false-positive predictions from the confusion matrix. Specifically, we conducted the classifier with best performance on the

whole balanced dataset created by oversampling the minority class. In result, we found 2,404 false-positive instances that might be potential off-label candidates, as shown in Table 2.

Table 2 Classification results of using *TD-PCL*

		Predicted	
		P	N
Actual	P	7098	812
	N	2404	5596

Since there have not been any medical databases that embody off-label usages, the most common and popular validation way is to find co-mention support from medical reports and literature [3]. Therefore, we examined the positive evidence for the detected 2,404 potential off-label drug uses in PubMed literature. PubMed is a publicly available medical repository that covers titles and abstracts of more than 26 million biomedical publications, managed by National Center for Biotechnology Information (NCBI). To validate the possibility that the findings might suggest an off-label relationship between the drug and disease, we searched their co-mentions in the abstracts in PubMed. By setting the threshold as ten, we found that 832 potential off-label drug-disease pairs were both mentioned in the abstract in at least ten articles.

V. CONCLUSION

Off-label drug uses are very common and inevitable to some extent in medical practice. Although gaining information about off-label drug uses could benefit the healthcare stakeholders including patients, physicians, biomedical scholars, pharmaceutical companies, and etc., there is no such data repository yet. This brings up the demand for a systematic way to detect off-label drug uses. Most of the off-label studies have exploited data resources such as EHR and clinical notes for the detection, however, we utilized social media data especially OHC data, with consideration of the increasing social media users and the large volume of valuable and timely user-generated contents. In addition, on the basis of our previous off-label research [9], we utilized tensor decomposition to deal with the sparsity and missing data problem in social media data.

In this work, we collected consumer-contributed contents from MedHelp and constructed a three-dimensional tensor (i.e. disease, drug, and user). We performed CP decomposition on the tensor which generated 100 resulting components. Then we adopted two methods to identify the off-label drug uses: (1) As each resulting component suggested a composition that included the drugs and diseases associated with the users and CP decomposition provided a way to evaluate the significance of those components, we utilized these advantages to rank the components and to find the significant drug-disease pairs that appeared on many users, thus identifying the off-label drug-disease relationships. (2) We utilized the advantage of tensor decomposition that it was able to capture the latent dependency and structure information of multidimensional dataset and then to reconstruct the three-dimensional dataset from that. We performed the heterogeneous network-based method proposed previously, specifically, meta-path-based method, on the reconstructed dataset to identify off-label drug uses. By ranking the decomposition results, we obtained some significant components and have done one case study on the most significant composition. The composition suggested relationships between two diseases and six drugs, and we found the medical evidence that explained 7 out of 12 co-occurrences though they are not originally indicated drug-disease pairs, which implied the reason and the possibility that the drugs were used in off-label way. By applying the heterogeneous network-based method on the reconstructed dataset from CP decomposition, we achieved better results than the method without using tensor decomposition, which indicated the effectiveness of tensor decomposition on handling missing data in multidimensional dataset and suggested a new way of using sparse social media data.

In summary, the experiment results have shown the advantage of tensor decomposition on social media data and off-label drug use detection, while there still exist several challenges that we will try to resolve in the future: firstly, tensor decomposition was used as a tool here to detect off-label drug use and the evaluation was conducted on the results of off-label detection, while in the future we could firstly evaluate the tensor composition algorithm in terms of its stability and computation cost. Secondly, we adopted CP decomposition because of its success in many areas especially in many healthcare topics, whereas we can compare its performance with other tensor decomposition algorithms. Thirdly, there usually involved several diseases and several drugs when identifying off-label uses from one significant resulting component, currently, we did case studies on all the possible relationships without considering the interactions between diseases and drugs, while we could find a way to analyze and filter the relationships first and then conduct the case study.

ACKNOWLEDGEMENT

This work was supported in part by the National Science Foundation under the Grant NSF-1741306, IIS-1650531, and DIBBs-1443019. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] Radley, D. C., Finkelstein, S. N., & Stafford, R. S. (2006). Off-label prescribing among office-based physicians. *Archives of internal medicine*, *166*(9), 1021-1026.

[2] Wittich, C. M., Burkle, C. M., & Lanier, W. L. (2012, October). Ten common questions (and their answers) about off-label drug use. In *Mayo Clinic Proceedings* (Vol. 87, No. 10, pp. 982-990). Elsevier.

[3] Jung, K., LePendu, P., Chen, W. S., Iyer, S. V., Readhead, B., Dudley, J. T., & Shah, N. H. (2014). Automated detection of off-label drug use. PloS one, 9(2), e89324.

[4] Yang, H., & Yang, C. C. (2016, October). Discovering Drug-Drug Interactions and Associated Adverse Drug Reactions with Triad Prediction in Heterogeneous Healthcare Networks. In *Healthcare Informatics (ICHI), 2016 IEEE International Conference on* (pp. 244-254). IEEE.

[5] Yang, C. C., Yang, H., Jiang, L., & Zhang, M. (2012, October). Social media mining for drug safety signal detection. In *Proceedings of the 2012 international workshop on Smart health and wellbeing* (pp. 33-40). ACM.

[6] Yang, C. C., Yang, H., & Jiang, L. (2014). Postmarketing drug safety surveillance using publicly available health-consumer-contributed content in social media. *ACM Transactions on Management Information Systems (TMIS)*, *5*(1), 2.

[7] Yang, H., & Yang, C. C. (2015). Using health-consumer-contributed data to detect adverse drug reactions by association mining with temporal analysis. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *6*(4), 55.

[8] Yang, C. C., & Zhao, M. (2017, August). Determining Associations with Word Embedding in Heterogeneous Network for Detecting Off-Label Drug Uses. In *Healthcare Informatics (ICHI), 2017 IEEE International Conference on*(pp. 496-501). IEEE.

[9] Zhao, M., & Yang, C. C. (2017, August). Automated Off-label Drug Use Detection from User Generated Content. In *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*(pp. 449-454). ACM.

[10] Zhao, M., & Yang, C. C. (2016, October). Mining online heterogeneous healthcare networks for drug repositioning. In *Healthcare Informatics (ICHI), 2016 IEEE International Conference on* (pp. 106-112). IEEE.

[11] Garg, L., Dauwels, J., Earnest, A., & Leong, K. P. (2014). Tensor-based methods for handling missing data in quality-of-life questionnaires. *IEEE journal of biomedical and health informatics*, *18*(5), 1571-1580.

[12] Turner, S. E. A. N., Nunn, A. J., Fielding, K., & Choonara, I. M. T. I. (1999). Adverse drug reactions to unlicensed and off-label drugs on paediatric wards: a prospective study. Acta Paediatrica, 88(9), 965-968.

[13] Neubert, A., Dormann, H., Weiss, J., Egger, T., Criegee-Rieck, M., Rascher, W., ... & Hinz, B. (2004). The impact of unlicensed and off-label drug use on adverse drug reactions in paediatric patients. Drug safety, 27(13), 1059-1067.

[14] Horen, B., Montastruc, J. L., & Lapeyre-Mestre, M. (2002). Adverse drug reactions and off-label drug use in paediatric outpatients. British journal of clinical pharmacology, 54(6), 665-670.

[15] Eguale, T., Buckeridge, D. L., Verma, A., Winslade, N. E., Benedetti, A., Hanley, J. A., & Tamblyn, R. (2016). Association of off-label drug use and adverse drug events in an adult population. JAMA internal medicine, 176(1), 55-63.

[16] Conroy, S., Choonara, I., Impicciatore, P., Mohn, A., Arnell, H., Rane, A., ... & Rocchi, F. (2000). Survey of unlicensed and off label drug use in paediatric wards in European countries. Bmj, 320(7227), 79-82.

[17] Jung, K., LePendu, P., & Shah, N. (2013). Automated detection of systematic off-label drug use in free text of electronic medical records. AMIA Summits on Translational Science Proceedings, 2013, 94.

[18] Mesgarpour, B., Müller, M., & Herkner, H. (2012). Search strategies to identify reports on "off-label" drug use in EMBASE. BMC medical research methodology, 12(1), 190.

[19] Xu, R., & Wang, Q. (2013). Large-scale extraction of accurate drug-disease treatment pairs from biomedical literature for drug repurposing. BMC bioinformatics, 14(1), 181.

[20] Chen, X., Liu, M. X., & Yan, G. Y. (2012). Drug–target interaction prediction by random walk on the heterogeneous network. Molecular BioSystems, 8(7), 1970-1978.

[21] Wang, W., Yang, S., Zhang, X., & Li, J. (2014). Drug repositioning by integrating target information through a heterogeneous network model. Bioinformatics, 30(20), 2923-2930.

[22] Martínez, V., Navarro, C., Cano, C., Fajardo, W., & Blanco, A. (2015). DrugNet: Network-based drug–disease prioritization by integrating heterogeneous data. Artificial intelligence in medicine, 63(1), 41-49.

[23] Huang, Y. F., Yeh, H. Y., & Soo, V. W. (2013). Inferring drug-disease associations from integration of chemical, genomic and phenotype data using network propagation. BMC medical genomics, 6(3), S4.

[24] Sillence, Elizabeth, Pam Briggs, Peter Richard Harris, and Lesley Fishwick. "How do patients evaluate and make use of online health information?." *Social science & medicine* 64, no. 9 (2007): 1853-1862.

[25] Powell, John, Nadia Inglis, Jennifer Ronnie, and Shirley Large. "The characteristics and motivations of online health information seekers: cross-sectional survey and qualitative interview study." *Journal of Medical Internet Research* 13, no. 1 (2011).

[26] Tustin, Nupur. "The role of patient satisfaction in online health information seeking." *Journal of health communication* 15, no. 1 (2010): 3-17.

[27] Bender, Jacqueline L., Maria-Carolina Jimenez-Marroquin, and Alejandro R. Jadad. "Seeking support on facebook: a content analysis of breast cancer groups." *Journal of medical Internet research* 13, no. 1 (2011).

[28] Dauwels, J., Garg, L., Earnest, A., & Pang, L. K. (2011, December). Handling missing data in medical questionnaires using tensor decompositions. In *Information, Communications and Signal Processing (ICICS) 2011 8th International Conference on* (pp. 1-5). IEEE.

[29] Acar, E., Aykut-Bingol, C., Bingol, H., Bro, R., & Yener, B. (2007). Multiway analysis of epilepsy tensors. *Bioinformatics*, *23*(13), i10-i18.

[30] De Vos, M., De Lathauwer, L., Vanrumste, B., Van Huffel, S., & Van Paesschen, W. (2007). Canonical decomposition of ictal scalp EEG and accurate source localisation: Principles and simulation study. *Computational intelligence and neuroscience*, *2007*.

[31] Yang, C. C., & Yang, H. (2015, November). Exploiting Social Media with Tensor Decomposition for Pharmacovigilance. In *Data Mining Workshop (ICDMW), 2015 IEEE International Conference on* (pp. 188-195). IEEE.

[32] Bader, B. W., & Kolda, T. G. (2007). *Tensor decompositions and their application* (No. SAND2007-4390C). Sandia National Laboratories (SNL-NM), Albuquerque, NM (United States); Sandia National Laboratories, Livermore, CA.

[33] D. M. Dunlavy, T. G. Kolda, and E. Acar, "Temporal link prediction using matrix and tensor factorizations," *ACM Transactions on Knowledge Discovery from Data (TKDD),* vol. 5, p. 10, 2011.

[34] J. D. Carroll and J.-J. Chang, "Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition," *Psychometrika,* vol. 35, pp. 283-319, 1970.

[35] R. A. Harshman, "Foundations of the parafac procedure: models and conditions for an" explanatory" multimodal factor analysis," 1970.

[36] Ho, J. C., Ghosh, J., Steinhubl, S. R., Stewart, W. F., Denny, J. C., Malin, B. A., & Sun, J. (2014). Limestone: High-throughput candidate phenotype generation via tensor factorization. *Journal of biomedical informatics*, *52*, 199-211.

[37] Song, Q., Ge, H., Caverlee, J., & Hu, X. (2017). Tensor Completion Algorithms in Big Data Analytics. *arXiv preprint arXiv:1711.10105*.

[38] Kossaifi, J., Panagakis, Y., & Pantic, M. (2016). Tensorly: Tensor learning in python. *arXiv preprint arXiv:1610.09555*.

[39] Simon, K. C., Chen, H., Schwarzschild, M., & Ascherio, A. (2007). Hypertension, hypercholesterolemia, diabetes, and risk of Parkinson disease. *Neurology*, *69*(17), 1688-1695.

[40] Gao, X., Chen, H., Schwarzschild, M. A., & Ascherio, A. (2011). Use of ibuprofen and risk of Parkinson disease. *Neurology*, *76*(10), 863-869.

[41] Onor, M. L., Trevisiol, M., & Aguglia, E. (2007). Rivastigmine in the treatment of Alzheimer's disease: an update. Clinical interventions in aging, 2(1), 17.

[42] Giofré, M. R., Meduri, G., Pallio, S., Calandra, S., Magnano, A., Niceforo, D., ... & Fries, W. (2000). Gastric permeability to sucrose is increased in portal hypertensive gastropathy. *European journal of gastroenterology & hepatology*, *12*(5), 529-533.

[43] Chen, X., Zhai, X., Kang, Z., & Sun, X. (2012). Lactulose: an effective preventive and therapeutic option for ischemic stroke by production of hydrogen. *Medical gas research*, *2*(1), 3.

[44] Ye Z, Rae CL, Nombela C, et al. Predicting beneficial effects of atomoxetine and citalopram on response inhibition in Parkinson's disease with clinical and neuroimaging measures. Human brain mapping. 2016 Mar 1;37(3):1026-37.

[45] Moosavi SM, Ahmadi M, Monajemi MB. Acute dystonia due to citalopram treatment: a case series. Global journal of health science. 2014 Nov;6(6):295.

[46] Fu, W., Ma, L., Zhao, X., Li, Y., Zhu, H., Yang, W., ... & Liu, H. (2015). Antidepressant medication can improve hypertension in elderly patients with depression. *Journal of Clinical Neuroscience*, *22*(12), 1911-1915.