Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)

Multitask Dyadic Prediction and Its Application in Prediction of Adverse Drug-Drug Interaction

Bo Jin,† Haoyu Yang,† Cao Xiao,* Ping Zhang,* Xiaopeng Wei,† Fei Wang¶
†Computer Science, Dalian University of Technology, Dalian, China 116024
*Health Analytics Research Group, IBM T.J.Watson Research Center, Yorktown Heights, NY 10598
¶ Healthcare Policy and Research, Weill Cornell Medical College, Cornell University, New York, NY 10065

Abstract

Adverse drug-drug interactions (DDIs) remain a leading cause of morbidity and mortality around the world. Identifying potential DDIs during the drug design process is critical in guiding targeted clinical drug safety testing. Although detection of adverse DDIs is conducted during Phase IV clinical trials, there are still a large number of new DDIs founded by accidents after the drugs were put on market. With the arrival of big data era, more and more pharmaceutical research and development data are becoming available, which provides an invaluable resource for digging insights that can potentially be leveraged in early prediction of DDIs. Many computational approaches have been proposed in recent years for DDI prediction. However, most of them focused on binary prediction (with or without DDI), despite the fact that each DDI is associated with a different type. Predicting the actual DDI type will help us better understand the DDI mechanism and identify proper ways to prevent it. In this paper, we formulate the DDI type prediction problem as a multitask dyadic regression problem, where the prediction of each specific DDI type is treated as a task. Compared with conventional matrix completion approaches which can only impute the missing entries in the DDI matrix, our approach can directly regress those dyadic relationships (DDIs) and thus can be extend to new drugs more easily. We developed an effective proximal gradient method to solve the problem. Evaluation on real world datasets is presented to demonstrate the effectiveness of the proposed approach.

1 Introduction

Drug-drug interaction (DDI) is a modification of the effect of a drug when administered with another drug, which is a common scenario for patients with complicated conditions such as cancer or other chronic diseases. Some DDIs could be an increase or a decrease in the effect, while some could be an adverse effect that even results in severe morbidity and mortality. Although detection of adverse DDIs is conducted during Phase IV clinical trials, there are still a large number of new DDIs founded by accidents after the drugs were put on market. These undetected adverse DDIs have become serious health threats and caused nearly $74,000$ emergency room visits and $195,000$ hospitalizations each year in the United States alone (Percha and Altman 2013). This brings an urgent need for methods that can identify DDIs earlier and more comprehensively.

With the arrival of the big data era, more and more healthcare related data are becoming readily available, so does the pharmaceutical industry. There are quite a few research works trying to leverage pharmaceutical research and development data in the task of DDI prediction. One of the prominent resource they are using is the chemical structure of the drugs, as many DDIs are essentially caused by the chemical-physical incompatibility of two drugs. For example, machine learning methods were developed for predicting DDIs by analyzing chemical structure similarity (Cheng and Zhao 2014b), implementing the chemical-protein interactome (Luo et al. 2014), modeling interaction profile fingerprints (Vilar et al. 2013), and exploiting pharmacointeraction network structure (Cami et al. 2013). There are also some efforts on predicting DDIs by integrating multiple molecular and pharmacological data (Yang, Xu, and Zeng 2014). The advantage of these methods lies in the fact that they rely mainly on chemical and bioactivity data from laboratory studies which could be obtained during preclinical phase rather than clinical records. As a result, they could potentially be used to predict DDIs years in advance, enabling drug safety professionals to better prioritize their limited investigative resources and take appropriate regulatory actions.

Despite their initial success, there are still limitations for existing computational approaches. For example,

- They only focus on binary predictions, which correspond to whether or not a DDI would happen. Actually there is a detailed type (e.g., hepatic failure, cough, dizziness, etc.) associated with each DDI. Obtaining the actual DDI types may help us understand better the mechanistic underlying the DDI and take proper preventative actions.
- They typically use drug features alone in the predictive model (e.g., logistic regression). However, in reality the interactions among different chemical compounds are also important factors that lead to DDIs.
- They usually work in an "imputation" manner, i.e., predicting potential interactions among existing drugs. It is also much demanding of predicting potential DDIs for new drugs.

To lift the aforementioned limitations, in this work, we

propose a multitask dyadic regression method that could simultaneously detect multiple potential types DDIs for a pair of drugs. Those drugs could be new or existing. Specifically, we treat the prediction of each specific DDI type is treated as a task and minimize the prediction loss over all different tasks jointly with proper regularizations. An effective proximal gradient method is developed to solve the optimization problem. We evaluate the proposed method on real world dataset. Results show that our model could not only accurately characterize the task relatedness and therefore significantly improve the prediction performance over baseline models, but also effectively control the confounding effects from covariates in observational clinical data and further enhance the predictions.

The rest of the paper is organized as follows. Section 2 introduces the building blocks of the proposed model. While in Section 3 we will review current works on adverse DDI detection. Next we will be ready to introduce the proposed model in Section 4 and evaluate its performance with real world data in Section 5. Finally we conclude our work and talk about future directions in Section 6.

2 Background

To prepare for the presentation of our method, in this section, we will briefly introduce the two main building blocks: dyadic prediction and multitask learning.

2.1 Dyadic Prediction

In dyadic prediction, we predict labels for pairs of objects, which can be considered as a matrix completion problem: we have partial observed label matrix $\mathcal{Y} = \{y_{i,j}\} \in \mathbb{R}^{|\mathcal{D}|\times|\mathcal{D}|}$ and dyads $\boldsymbol{d}_i, \boldsymbol{d}_j \in \mathcal{D}$, where $\boldsymbol{d}_i$ and $\boldsymbol{d}_j$ impact the row and column of $\mathcal{Y}$, respectively. Let $\{(\boldsymbol{d}_i, \boldsymbol{d}_j), y_{i,j}\}$ be the training set, and $\{(\tilde{\boldsymbol{d}}_i, \tilde{\boldsymbol{d}}_j), \tilde{y}_{i,j}\}$ be the unobserved target set, dyadic prediction is to predict $\tilde{y}_{i,j}$ from $(\tilde{\boldsymbol{d}}_i, \tilde{\boldsymbol{d}}_j)$ given the model trained from training set.

Traditional dyadic prediction methods often suffer from the cold-start problem. If one of $\tilde{\boldsymbol{d}}_i$ or $\tilde{\boldsymbol{d}}_j$ is not present in the training set, then we cannot directly compute the label $\tilde{y}_{i,j}$. Intuitively, if the entire row or column in $\mathcal{Y}$ are not observed, we cannot predict the value in this row or column.

In this work, we take an inductive approach to train a model that could characterize the structural similarity between any entity pairs $\boldsymbol{d}_i$ and $\boldsymbol{d}_j$ in the training set, represented via a feature vector. The proposed method addresses the "cold-start" issue and will be able to estimate the probability of DDI events between the unobserved drugs and other drugs only based on the structural information and the structural similarity vectors from the training set. Therefore, the proposed method fits the need of predicting potential DDIs in the preclinical phase.

2.2 Multitask Learning

Multitask learning is a general approach that incorporates task relatedness during learning. In many real-life scenarios, subjects can relate to each other in some way. Multitask learning method is such an inductive transfer learning approach that improves generalization by exploiting the connections amongst related subjects as an inductive bias and has been proved by literature to have better generalization capabilities (Ando and Zhang 2005). In learning, multitask learning learns tasks in parallel while using a shared prior representation that encodes the "relatedness", thus what is learned for each task can help other tasks be learned better (Caruana 1998). In this work, we formulate the prediction of each DDI event as one dyadic prediction task, and solve these related DDI tasks in a multitask setting and achieve better predictive power.

3 Related Work

A few computational and mathematical modeling methods have been proposed to study the interactions between drugs, most of which rely on the understanding of the mechanisms underlying each interaction types (Jin et al. 2011). In contrast to computational modeling, recently more focus has been on predictive modeling approach. Much related works have developed ways of computing similarity scores between drugs or pairs of drugs to be used as features for machine learning classifiers (Cheng and Zhao 2014a; Gottlieb et al. 2012; Zhang et al. 2015). Sophisticated algorithms such as restricted Boltzmann machines and matrix factorization are especially effective in combining two types of similarities by learning latent representation (Cao et al. 2015; Wang and Zeng 2013). However, they do not inherently support inductive prediction.

Methodology-wise, the proposed method is related to Inductive Matrix Completion (Natarajan and Dhillon 2014) which makes a low-rank assumption on the coefficient matrix for capturing dyadic feature interactions. Factorization Machine (FM) (Rendle 2010) is another related model, which predicts the preference score with a function including both linear and nonlinear feature components, where the nonlinear components capture the pairwise or high order feature interactions. The difference between our model and FM is that FM is still a regular single-input regression model, while our model predicts the dyadic relationships among pairs of data entities. One interesting observation though, is if we make all the input data entities equal, i.e., we learn the high-order "relationship" between the data entity and itself, then we can recover FM if we impose low-rank assumption on the coefficients for the high-order feature interactions.

In addition, Multi-view Machines (Cao et al. 2016). Multi-view Machines generalize the FM model to capture all higher-order interactions between different data views and *jointly* factorize all of them via the CP tensor decomposition (Carroll and Chang 1970; Harshman 1970b). On one hand, this decision limits the number of parameters to learn. However, it restricts to a target model where all the low-rank factors are shared between parameters reflecting interactions of different order. Our model is more flexible, allowing for both shared and non-shared low-rank factors to be learned.

There are works that are limited to modeling the relationship between two data domains, such as the Sparse Factorization Machine (Xu et al. 2016), the Conditional High-

Order Boltzmann Machine (Huang, Wang, and Wang 2015) and the ConsMRF (Drumond, Diaz-Aviles, and Schmidt-Thieme 2016). Also, the Hierarchical Interaction Representation in (Liu, Wu, and Wang 2015) can model multi-entity interactions, but assumes a certain interacting order of various data domains, which has to be manually determined and is application-dependent. For example in latent collaborative retrieval, they assume that only the joint interactions between users and queries with documents have to be taken into account. In contrast, the proposed model can allow any interaction between data domains, so that the important ones are automatically learned based on the input data.

4 Method

4.1 The Proposed Model

In this section, we introduce the proposed multitask dyadic prediction method as well as describe its technical details. Our model is motivated by the real world problems of DDI prediction. Denote $\mathbf{S}_k$ as the coefficient matrix capturing the relationship between i-th drug and j-th drug with respect to k-th DDI event, we could use dyadic prediction as in Formula 1 to characterize such a specific DDI event.

$$f_k(\boldsymbol{d}_i, \boldsymbol{d}_j) = \boldsymbol{d}_i^\top \mathbf{S}_k \boldsymbol{d}_j + b \tag{1}$$

If there are V individual DDIs to be predict, we will need to learn V different coefficient matrices $\mathbf{S}$. However, these DDI tasks are often related. For example, "Heart Rate Increased (HRI)" is often associated with "Body Temperature Increased (BTI)", if a specific drug pair $\boldsymbol{d}_i, \boldsymbol{d}_j$ causes HRI, then it is very likely they will cause BTI, too. Due to such relatedness, rather than learning the tasks (e.g. DDI events) in isolation and ignoring their relatedness, it would be beneficial for us to exploit the connections amongst them by learning the tasks simultaneously. Therefore, multitask learning method becomes a natural setting in terms of leveraging the shared information contained in the variables to improve generalization power.

To solve the dyadic prediction problem, we optimize the following objective function as shown in Formula 2.

$$\mathcal{L}_k = \frac{1}{N} \sum_{(i,j)\in N} \ell(f_k(\boldsymbol{d}_i, \boldsymbol{d}_j), \mathcal{Y}_k(\boldsymbol{d}_i, \boldsymbol{d}_j)) + \lambda_k \Omega(\mathbf{S}_k) \tag{2}$$

where $\ell(f_k(\boldsymbol{d}_i, \boldsymbol{d}_j), \mathcal{Y}_k(\boldsymbol{d}_i, \boldsymbol{d}_j))$ is the regression loss of function f, Ω is an aggregation of the penalties imposed on the model parameters. The coefficient matrices $\mathbf{S}_k$ in Formula 1 can be optimized by minimize function $\mathcal{L}_k$. The loss function ℓ in the formula is determined by the type of target value $\mathcal{Y}_k(\boldsymbol{d}_i, \boldsymbol{d}_j)$. And we could solve the parameters jointly by minimizing the objective function in Formula 3.

$$\mathcal{L} = \frac{1}{N} \sum_{(i,j)\in N} \sum_{k\in V} \ell(f_k(\boldsymbol{d}_i, \boldsymbol{d}_j), \mathcal{Y}_k(\boldsymbol{d}_i, \boldsymbol{d}_j)) + \sum_{k\in V} \lambda_k \Omega(\mathbf{S}_k) \tag{3}$$

In Formula 3, although we solve different coefficient matrices together, there is still no inner relationship among different coefficient matrices. However, with multitask learning approach, we use a coefficient tensor $\mathcal{S}$ to capture the dyadic relation among two drugs and DDI events rather than a matrix. In this way, the relationships among different DDI events will be captured in the additional dimension of the tensor.

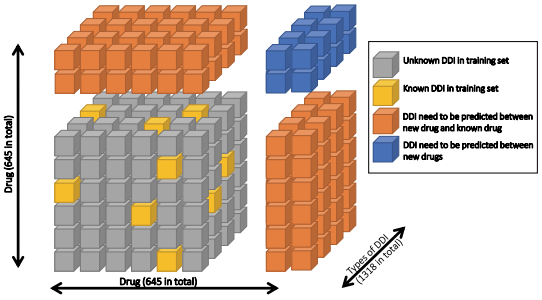

Figure 1: Illustration of the Drug-DDI Tensor

4.2 Tensorial Interpretation

The multitask dyadic model can effectively be formulated as a tensor (see Figure 1 for an illustration). In the figure, each cube represent the value of $\mathcal{Y}_{i,j,k}$, where the gray cube for 0 and the yellow cube for 1.

Now we formalize its tensorial notation and formulation. Denote $\mathcal{D} \in \mathbb{R}^{k\times m}$ as the data collection of drug's structure information, where k refers to the number of drugs and m is the dimension of drug's feature vector. We also denote $\boldsymbol{d}_i$ as the feature vector for the i-th drug in $\mathcal{D}$. And represent the data collection of DDI events with $\mathcal{E} \in \mathbb{R}^{v\times v}$, where $\boldsymbol{e}_j$ is the j-th DDI event in $\mathcal{E}$ with dimensionality $\boldsymbol{d}^v$. $\mathcal{Y} \in \mathbb{R}^{k\times k\times v}$ is the label tensor, in which $\mathcal{Y}_{i,j,k}$ is the relationship value among the i-th and j-th data object in $\mathcal{D}$ and the k-th data object in $\mathcal{E}$. Given these notation, our goal is to learn a function as in Formula 4 that characterizes dyadic DDI relationship $f(\boldsymbol{d}_i, \boldsymbol{d}_j, \boldsymbol{e}_k)$ to predict $\mathcal{Y}_{i,j,k}$.

$$f(\boldsymbol{d}_i, \boldsymbol{d}_j, \boldsymbol{e}_k) = \mathcal{S} \times_1 \boldsymbol{d}_i \times_2 \boldsymbol{d}_j \times_3 \boldsymbol{e}_k + b \tag{4}$$

where $\mathcal{S} \in \mathbb{R}^{m\times m\times v}$ becomes the coefficient tensor instead of the matrix in dyadic prediction, while $\times_k$ is the mode-k product.

For notational convenience, we use $f_{i,j,k}$ to represent the model function value of $f(\boldsymbol{d}_i, \boldsymbol{d}_j, \boldsymbol{e}_k)$, and $\mathcal{Y}_{i,j,k}$ is the ground truth relationship among the data entities of drugs and DDIs (e.g., $\mathcal{Y}_{i,j,k} = 1$ if the i-th and the j-th drug would lead to the k-th DDI, $\mathcal{Y}_{i,j,k} = 0$ otherwise).

Similar to Formula 3, we can solve all the models jointly by minimizing the following objective:

$$\mathcal{L} = \frac{1}{N} \sum_{(i,j,k)\in N} \ell(f_{i,j,k}, \mathcal{Y}_{i,j,k}) + \lambda \Omega(\mathcal{S}) \tag{5}$$

Loss Function In the objective function above, the loss function $\ell(f_{i,j,k}, \mathcal{Y}_{i,j,k})$ depends on the type of $\mathcal{Y}_{i,j,k}$. For binary $\mathcal{Y}$, we use logistic loss or hinge loss; while for continuous $\mathcal{Y}$, we choose square loss. In our case, we employ

logistic loss as in Formula 6 since $\mathcal{Y}_{i,j,k}$ is binary.

$$\ell(f_{i,j,k}, \mathcal{Y}_{i,j,k}) = -(1-\mathcal{Y}_{i,j,k}) * log(\frac{\boldsymbol{e}^{-f_{i,j,k}}}{1+\boldsymbol{e}^{-f_{i,j,k}}}) - \mathcal{Y}_{i,j,k} * log(\frac{1}{1+\boldsymbol{e}^{-f_{i,j,k}}}) \quad (6)$$

Optimization For regularization of the model parameters, we use proximal gradient methods to solve the following optimization problem:

$$\min_{\boldsymbol{u}\in\mathcal{H}} \mathcal{J}(\boldsymbol{u}) + \mathcal{R}(\boldsymbol{u}) \quad (7)$$

where $\mathcal{J}$ is a convex and differentiable function with Lipschitz continuous gradient, $\mathcal{R}$ is a convex and lower semi-continuous function which is possibly nondifferentiable, and $\mathcal{H}$ is a set, typically a Hilbert space. In proximal methods, the usual criterion that u minimizes $\mathcal{J}(\boldsymbol{u}) + \mathcal{R}(\boldsymbol{u})$ if and only if $\nabla_{\boldsymbol{u}}(\mathcal{J}(\boldsymbol{u}) + \mathcal{R}(\boldsymbol{u})) = 0$ is replaced by $0 \in \partial_{\boldsymbol{u}}(\mathcal{J}(\boldsymbol{u}) + \mathcal{R}(\boldsymbol{u}))$, where ∂ is the subdifferential operator.

One key operator we need to define for proximal methods is the proximal operator. Given a convex function $\psi : \mathcal{H} \to \mathbb{R}$, we can define its proximal operator $\text{prox}_{\psi} : \mathcal{H} \to \mathcal{H}$ as in Formula 8.

$$\text{prox}_{\psi}(\boldsymbol{z}) = \arg\min_{\boldsymbol{u}\in\mathcal{H}} \psi(\boldsymbol{u}) + \frac{1}{2}\|\boldsymbol{u}-\boldsymbol{z}\|_2^2 \quad (8)$$

The proximity operator can be seen as a generalization of a projection. And u^* is the optimal solution to the problem of Formula (7) if and only if Formula 9 holds.

$$u^* = \text{prox}_{\gamma\mathcal{R}}\left(u^* - \gamma\nabla\mathcal{J}(u^*)\right) \quad (9)$$

where $\gamma > 0$ is a constant.

There are alternative proximal operators which can be used for generalization, such as ℓ_1 norm, ℓ_2 norm, elastic net, and group sparsity. In this case, $\mathcal{R}(\boldsymbol{u}) = \frac{1}{2}\|\boldsymbol{u}\|_2^2$, we employ ℓ_2 norm as proximal operator:

$$\text{prox}_{\gamma\mathcal{R}}(\boldsymbol{u})\big|_i = \frac{1}{1+\gamma}u_i \quad (10)$$

Low-Rank Case In this paper, we assume that the high-order coefficient tensors are low rank, i.e., they can be approximated by the product of a set of low-rank matrices. For order-3 relationship coefficient matrices, we can adopt low-rank CANDECOMP/PARAFAC(CP) decomposition(Harshman 1970a), i.e.,

$$\mathcal{S}^{mmv} \approx \mathbf{O}\otimes\mathbf{P}\otimes\mathbf{Q} \quad (11)$$

where $\mathcal{S} \in \mathbb{R}^{m\times m\times v}$ is the low-rank tensor. $\mathbf{O}$, $\mathbf{P}$, and $\mathbf{Q}$ denote the three matrices obtained from the CP decomposition of low-rank tensor $\mathcal{S}$. $\mathbf{O} \in \mathbb{R}^{m\times r}$ with $r \ll m$, $\mathbf{P} \in \mathbb{R}^{m\times r}$, and $\mathbf{Q} \in \mathbb{R}^{v\times r}$. The regularization we mentioned in the prior section can be added to the factorized matrices to impose different kinds of requirements.

Challenges in Solving Let $\mathbf{O}$ corresponding to the drug $\boldsymbol{d}_i$, $\mathbf{P}$ corresponding to the drug $\boldsymbol{d}_j$, $\mathbf{Q}$ corresponding to the drug $\boldsymbol{e}_k$, the optimal problem we want to solve becomes Formula 12.

$$\min_{\mathbf{O},\mathbf{P},\mathbf{Q}} \mathcal{J} + \lambda_{\mathbf{O}}\Omega(\mathbf{O}) + \lambda_{\mathbf{P}}\Omega(\mathbf{P}) + \lambda_{\mathbf{Q}}\Omega(\mathbf{Q}) \quad (12)$$

where we have

$$\mathcal{J} = -\frac{1}{|N|}\sum_{(i,j,k)\in N} y_{i,j,k} * log(\frac{1}{1+\boldsymbol{e}^{-f_{i,j,k}}}) + (1-y_{i,j,k}) * log(1-\frac{1}{1+\boldsymbol{e}^{-f_{i,j,k}}})$$

And then Formula 4 can be reformulated as follows.

$$f_{i,j,k} = (\mathbf{O}\otimes\mathbf{P}\otimes\mathbf{Q}) \times_1 \boldsymbol{d}_i \times_2 \boldsymbol{d}_j \times_3 \boldsymbol{e}_k + b \quad (13)$$

There are three variables in the target function, we employ alternating proximal gradient descent (PGD) method to solve all the parameters in the model. Algorithm [1] is the Alternating PGD to solve our optimal problem of Formula (12). The algorithm mainly consists of three steps. We update one variable in each of the three steps alternatively.

Algorithm 1 Alternating PGD algorithm

```
Initialize O(0), P(0), Q(0)
for t = 1, 2, ··· , T do
    O(t+1) = APGD_O(O(t), P(t), Q(t))
    P(t+1) = APGD_P(O(t+1), P(t), Q(t))
    Q(t+1) = APGD_Q(O(t+1), P(t+1), Q(t))
end for(while converge)
return O(T), P(T), Q(T)
```

However, traditional PGD algorithm converges slow, which is not a viable solution in our case due to high data dimensions.

Solving with Accelerated Proximal Gradient To solve such a challenge, we employ accelerated PGD algorithm instead. The accelerated version of the basic PGD algorithm takes the former iterations result into consideration, which can be interpreted as momentum method. In order to further improve the convergence rate, an adaptive restart method(O'Donoghue and Candès 2015) is employed. This method takes the current iteration as the new starting point, and resets the related parameter to the initial value.

As we should optimize the three variables in turn, that is, we apply three proximal gradient descent (PGD) algorithms for the three variables. The PGD Algorithms for O, P, and Q have the same structures. Algorithm [2] is the accelerated PGD algorithm for optimizing the parameter of $\mathbf{O}$.

The derivative of function $\mathcal{J}$ in Algorithm [2] is as following:

$$\frac{\partial\mathcal{J}}{\partial\mathbf{O}} = \frac{1}{|N|}\sum_{(i,j,k)\in N}(\frac{1}{1+\boldsymbol{e}^{-f_{i,j,k}}} - y_{i,j,k}) * \frac{\partial f_{i,j,k}}{\partial\mathbf{O}} = \frac{1}{|N|}\sum_{(i,j,k)\in N}(\frac{1}{1+\boldsymbol{e}^{-f_{i,j,k}}} - y_{i,j,k}) * \boldsymbol{d}_i\boldsymbol{d}_j^\top\mathbf{P}diag(\boldsymbol{e}_k^\top\mathbf{Q})$$

$$\frac{\partial\mathcal{J}}{\partial\mathbf{P}} = \frac{1}{|N|}\sum_{(i,j,k)\in N}(\frac{1}{1+\boldsymbol{e}^{-f_{i,j,k}}} - y_{i,j,k}) * \frac{\partial f_{i,j,k}}{\partial\mathbf{P}} = \frac{1}{|N|}\sum_{(i,j,k)\in N}(\frac{1}{1+\boldsymbol{e}^{-f_{i,j,k}}} - y_{i,j,k}) * \boldsymbol{d}_j\boldsymbol{d}_i^\top\mathbf{O}diag(\boldsymbol{e}_k^\top\mathbf{Q})$$

Algorithm 2 Accelerated PGD for O with fixed P and Q

Initialize $\beta \in (0,1)$ $\gamma_O \in (0,1)$, $\lambda_O \in (0,1)$, $\boldsymbol{z}_O = \mathbf{O}, \mathbf{O}_{old} = \mathbf{O}, \theta_O = 1$
for $t = 1, 2, \cdots, T$ **do**
  Let $\gamma_O = \gamma_O(t-1)$
  while True **do**
    $\boldsymbol{z}_O = \text{prox}_{\lambda_O \text{R}}(\mathbf{O}(t) - \gamma_O \frac{\partial \mathcal{J}}{\partial \mathbf{O}}(\mathbf{O}(t)))$
    $\hat{\mathcal{J}}(\boldsymbol{z}_O) = \mathcal{J}(\mathbf{O}(t)) + trace((\frac{\partial \mathcal{J}}{\partial \mathbf{O}})^{\text{T}}(\boldsymbol{z}_O - \mathbf{O}(t))) + \frac{1}{2\gamma_O}\|\boldsymbol{z}_O - \mathbf{O}(t)\|_2^2$
    if $\mathcal{J}(\boldsymbol{z}_O) \leq \hat{\mathcal{J}}(\boldsymbol{z}_O)$ **then**
      break
    end if
    $\gamma_O = \beta \gamma_O$
  end while
  $\theta_O = \frac{2}{1+\sqrt{1+\frac{4}{\theta_O^2}}}$
  if $trace((\mathbf{O} - \boldsymbol{z}_O)^{\text{T}} * (\boldsymbol{z}_O - \mathbf{O}_{old})) > 0$ **then**
    $\mathbf{O}(t+1) = \mathbf{O}_{old}$
    $\theta_O = 1, \gamma_O = 1$
  else
    $\mathbf{O}(t+1) = \boldsymbol{z}_O + (1-\theta_O) * (\boldsymbol{z}_O - \mathbf{O}_{old})$
  end if
  $\gamma_O(t) = \gamma_O$
  $\mathbf{O}_{old} = \boldsymbol{z}_O$
end for(while converge)
return $\mathbf{O}(T)$

$$\frac{\partial \mathcal{J}}{\partial \mathbf{Q}} = \frac{1}{|N|} \sum_{(i,j,k) \in N} (\frac{1}{1+\boldsymbol{e}^{-f_{i,j,k}}} - y_{i,j,k}) * \frac{\partial f_{i,j,k}}{\partial \mathbf{Q}}$$
$$= \frac{1}{|N|} \sum_{(i,j,k) \in N} (\frac{1}{1+\boldsymbol{e}^{-f_{i,j,k}}} - y_{a,b,c}) * \boldsymbol{e}_k \boldsymbol{d}_j^{\top} \mathbf{P} diag(\boldsymbol{d}_i^{\top} \mathbf{O})$$

5 Experiments

In this section, we evaluate the effectiveness of our model with experiment on real world data and discuss the results.

5.1 Datasets

The DDI data we used in evaluation is extracted from FDA Adverse Event Reporting System (FAERS)[1]. The FAERS contains information on adverse events submitted to FDA, which is designed to support FDA's post-marketing safety surveillance program for drugs and therapeutic biological products. Mined from FAERS, the Twosides database[2] (Tatonetti et al. 2012) is a resource of polypharmacy side effects for pairs of drugs. It contains only side effects caused by the combination of drugs rather than by any single drug. In this study, we only used the unsafe co-prescriptions from Twosides database as known set of DDIs. There are 645 drugs and 1318 DDI events, in which $59,220$ distinct pairs of drugs associated with DDI reports.

We used PubChem substructure fingerprint[3] to construct drug features. Each drug was represented by an 881-dimensional binary profile whose elements encode the presence or absence of each substructure by 1 or 0, respectively.

[1]https://open.fda.gov/data/faers/

[2]http://tatonettilab.org/resources/tatonetti-stm.html

[3]http://tinyurl.com/y7svnm

5.2 Evaluations

To ensure the validity of the test cases, the validation was carried out by holding out all the DDIs associated with a fixed percentage of the drugs, rather than holding out DDIs directly. To be specific, we randomly selected a fixed percentage (10%) of drugs for testing, and considered all DDIs associated with these drugs as testing set. Then we constructed the models with the remaining DDIs as the training set. The model parameters were tuned with cross validation based on the training set. Models were tested on the testing set only after all model parameter tuning has been done.

We use receiver operator characteristic (ROC) curves and precision-recall (PR) curves to evaluate the proposed method. In the ROC and PR analytics, we utilized DDI interactions from Twosides database as positive samples, and the complement set of Twosides DDI interactions as negative samples. Here since the number of positive samples in the test data set is far less than that of the negative samples, the PR curve can describe the predicting results better.

5.3 Experiment Results

To evaluate its performance, we compared the proposed model with several state-of-the-art alternatives as baselines, including Logistic Regression(LR), Factorization Machine(FM), and Support Vector Machine(SVM). We performed the following experiments. First, we try to predict 2 randomly selected DDIs on the test set. As shown in Figure 2, the proposed model outperforms all models in terms of both ROC and PR curves. Particularly for the PR curve, ours learns 64.2% better than LR, which is the best among all baselines. For the ROC curve as shown in Figure 3, ours outperforms LR by 15.6%, while the other baselines including SVM and FM barely have any predictive power. Second, we compare based on top 50 frequent DDIs from the Twosides database. Figure 4 and 5 show the predicting result for the 50 DDIs. Results are consistent with previous experiments on 2 DDIs. We can observe that our model consistently gains the best performance, while the baseline models still have very low predictive ability.

5.4 Case Study

Here we also present case study to visualize the effectiveness of the proposed model using drug-drug networks. First, we construct the drug-drug networks that indicate whether any two drugs would result in a specific DDI. As shown in Figure 6, the node in the network denotes a drug. The ID of the drug is shown on the node. The edge between the two nodes denotes the existence of DDI. It is easier to understand that some specific drugs would have a higher risk to have DDI than other drugs. In the network, the size of the node denotes the degree of risk of a drug. We classify the degree of risk into different levels (in different colors), e.g. high-risk (blue), and low-risk (white). The red nodes denote the forecasting errors of drugs.

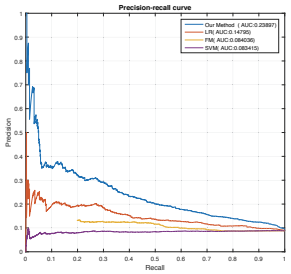

Figure 2: PR curve (2 DDIs)

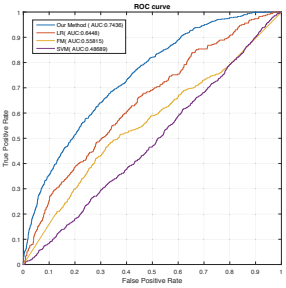

Figure 3: ROC curve (2 DDIs).

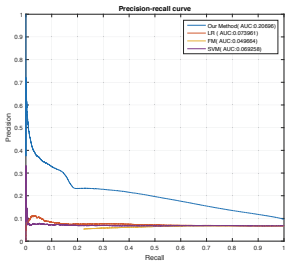

Figure 4: PR curve (50 DDIs)

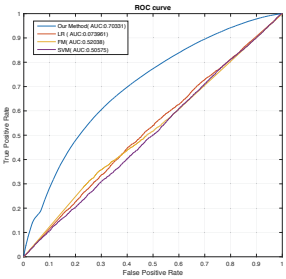

Figure 5: ROC curve (50 DDIs)

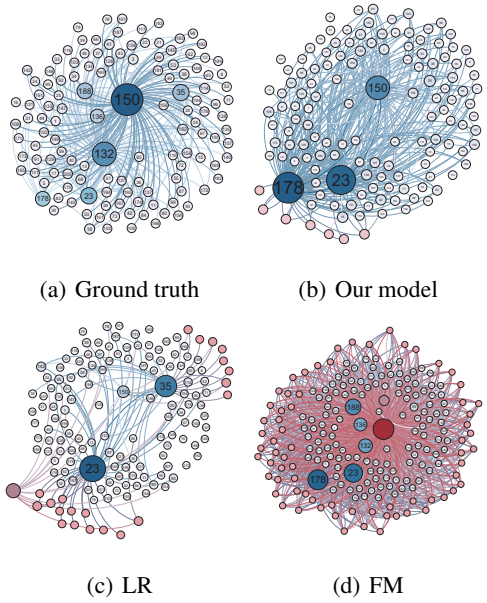

Figure 6: Comparison of connection characterization

As shown in Figure 6(a), Drugs with ID #150, #132, #188, #23, #35, #178, #136 are such high-risk drugs according to ground truth. Figure 6(b) shows that the proposed model correctly predicts three drugs (#150, #178 and #23) as high-risk ones. While the results of LR, as shown in Figure 6(c), is sparser than the ground truth, which indicates lots of existence drug-drug interactions are missed in the prediction of LR. In addition, although the number of predicted high-risk drugs (No. 23, 35, 150) of LR model is same with our model, the forecasting errors of LR model is much higher. As for FM model (in Figure 6(d)), although it can predict almost all the high-risk drugs, it also makes more incorrect prediction

5.5 False Positive Analysis

Since the ground truth data used in this study come from user report systems, it cannot cover all existing DDI cases, thus reviewing false positive predictions could potentially introduce novel knowledge about DDIs and bring huge value to pharmacovigilance studies.

One pair of drugs we discover are "Amiodarone (an antiarrhythmic medication used to treat and prevent a number of types of irregular heartbeats) " and "Buspirone (an antianxiety drug)". Although there is no record in FAERS TWOSIDES data showing they could cause DDI, the proposed model predicts a few their DDIs with high likelihood (in parenthesis), including severe DDIs such as anaemia (0.8301) neumonia (0.8088), asthenia (0.8011), arterial pressure NOS decreased (0.7700), acute kidney failure (0.6945), sepsis (0.6903), kidney failure (0.6834), and AFIB (0.6778). Such discovery could be well explained from the chemical structures of the drugs: Amiodarone is a CYP3A4 inhibitors, while Buspirone needs CYP3A4 for metabolism. Taking them together would greatly lower the metabolic rate of Buspirone and cause Buspirone to be highly concentrated in blood. Consequently many of the predicted DDIs would occur and become huge risk factors to patients.

6 Conclusion

In summary, we presented a new multitask dyadic prediction model to predict adverse drug-drug interactions (DDIs). Our approach direct regresses those dyadic relationships (DDIs) and thus can be extend to new drugs more easily. We further developed an effective proximal gradient method to solve the problem. We evaluated the performance of the proposed method in a large real world clinical observational databases (Twosides) and also demonstrated the efficacy and utility of the proposed method with case studies.

Future direction could include better models for more accurate DDIs identification, as well as a more efficient solving algorithm that could support large scale DDI detection. This work essentially establish a extendable foundation for us to pursue these future directions.

7 Acknowledgment

Fei Wang is partially supported by National Science Foundation under Grant Number IIS-1650723.

References

Ando, R. K., and Zhang, T. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research* 6(Nov):1817–1853.

Cami, A.; Manzi, S.; Arnold, A.; and Reis, B. 2013. Pharmacointeraction network models predict unknown drug-drug interactions. *PLoS ONE* 8(4).

Cao, D.-S.; Xiao, N.; Li, Y.-J.; Zeng, W.-B.; Liang, Y.-Z.; Lu, A.-P.; Xu, Q.-S.; and Chen, A. 2015. Integrating multiple evidence sources to predict adverse drug reactions based on a systems pharmacology model. *CPT: Pharmacometrics & Systems Pharmacology* 4(9):498–506.

Cao, B.; Zhou, H.; Li, G.; and Yu, P. S. 2016. Multi-view machines. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, WSDM '16, 427–436. New York, NY, USA: ACM.

Carroll, J. D., and Chang, J.-J. 1970. Analysis of individual differences in multidimensional scaling via an n-way generalization of "eckart-young" decomposition. *Psychometrika* 35(3):283–319.

Caruana, R. 1998. *Multitask Learning*. Boston, MA: Springer US. 95–133.

Cheng, F., and Zhao, Z. 2014a. Machine learning-based prediction of drug–drug interactions by integrating drug phenotypic, therapeutic, chemical, and genomic properties. *Journal of the American Medical Informatics Association* 21(e2):e278–e286.

Cheng, F., and Zhao, Z. 2014b. Machine learning-based prediction of drug–drug interactions by integrating drug phenotypic, therapeutic, chemical, and genomic properties. *Journal of the American Medical Informatics Association* 21(e2).

Drumond, L.; Diaz-Aviles, E.; and Schmidt-Thieme, L. 2016. Multi-Relational Learning at Scale with ADMM. *ArXiv e-prints*.

Gottlieb, A.; Stein, G. Y.; Oron, Y.; Ruppin, E.; and Sharan, R. 2012. Indi: a computational framework for inferring drug interactions and their associated recommendations. *Molecular Systems Biology* 8(1).

Harshman, R. 1970a. Foundations of the parafac procedure: Models and conditions for an explanatory multimodal factor analysis. *UCLA Working Papers in Phonetics* 16.

Harshman, R. A. 1970b. Foundations of the parafac procedure: Models and conditions for an "explanatory" multimodal factor analysis. *UCLA Working Papers in Phonetics* 16:1–84.

Huang, Y.; Wang, W.; and Wang, L. 2015. Conditional high-order boltzmann machine: A supervised learning model for relation learning. In *The IEEE International Conference on Computer Vision (ICCV)*.

Jin, G.; Zhao, H.; Zhou, X.; and Wong, S. T. C. 2011. An enhanced petri-net model to predict synergistic effects of pairwise drug combinations from gene microarray data. *Bioinformatics* 27(13):i310–i316.

Liu, Q.; Wu, S.; and Wang, L. 2015. Collaborative prediction for multi-entity interaction with hierarchical representation. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, 613–622. New York, NY, USA: ACM.

Luo, H.; Zhang, P.; Huang, H.; Huang, J.; Kao, E.; Shi, L.; He, L.; and Yang, L. 2014. Ddi-cpi, a server that predicts drug-drug interactions through implementing the chemical-protein interactome. *Nucleic Acids Research* 42(7):W46–52.

Natarajan, N., and Dhillon, I. S. 2014. Inductive matrix completion for predicting gene–disease associations. *Bioinformatics* 30(12):i60–i68.

O'Donoghue, B., and Candès, E. 2015. Adaptive restart for accelerated gradient schemes. *Foundations of Computational Mathematics* 15(3):715–732.

Percha, B., and Altman, R. B. 2013. Informatics confronts drug–drug interactions. *Trends in Pharmacological Sciences* 34(3):178 – 184.

Rendle, S. 2010. Factorization machines. In *2010 IEEE International Conference on Data Mining, ICDM'10*, 995–1000.

Tatonetti, N. P.; Ye, P. P.; Daneshjou, R.; and Altman, R. B. 2012. Data-driven prediction of drug effects and interactions. *Science Translational Medicine* 4(125):125ra31–125ra31.

Vilar, S.; Uriarte, E.; Santana, L.; Tatonetti, N. P.; and Friedman, C. 2013. Detection of drug-drug interactions by modeling interaction profile fingerprints. *PLoS ONE* 8(3):1–11.

Wang, Y., and Zeng, J. 2013. Predicting drug-target interactions using restricted boltzmann machines. *Bioinformatics* 29(13):i126–i134.

Xu, J.; Lin, K.; Tan, P.-N.; and Zhou, J. 2016. Synergies that matter: Efficient interaction selection via sparse factorization machine. In *Proceedings of the 2016 SIAM International Conference on Data Mining*, 108–116.

Yang, F.; Xu, J.; and Zeng, J. 2014. Drug-target interaction prediction by integrating chemical, genomic, functional and pharmacological data. *Pac Symp Biocomput.*

Zhang, P.; Wang, F.; Hu, J.; and Sorrentino, R. 2015. Label propagation prediction of drug-drug interactions based on clinical side effects. *Scientific reports* 5.