Drug Target Interaction Prediction with Non-random Missing Labels

Anonymous Author(s)

ABSTRACT

We introduce a novel probabilistic model, **F**actorization with **N**on-random **M**issing **L**abels (FNML), for **D**rug-**T**arget **I**nteraction (DTI) prediction. DTI prediction is usually modeled as a binary classification problem. Unlike previous studies which label unknown DTIs as negative samples, we assume the unknown DTIs are labels that are missing not at random. For example, negative DTI labels are more likely to be missing because biomedical researchers priorize to study DTIS that are more likely to be positive. The proposed model, FNML models the generative process for the DTI labels (i.e. the labels are positive or negative) and responses (i.e. the labels are observed or missing). In particular, the probability of observing or missing a label is associated with the sign of the label. Experimental results show that FNML outperforms state-of-the-art methods.

KEYWORDS

Missing Not At Random, Drug Target Interaction Prediction, Probabilistic Factor Models

ACM Reference format:
Anonymous Author(s). 2016. Drug Target Interaction Prediction with Non-random Missing Labels. In *Proceedings of ACM Conference, Washington, DC, USA, July 2017 (Conference'17),* 4 pages.
DOI: 10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Drug-**T**arget **I**nteraction (DTI) is fundamental to drug discovery and design. As biochemical experimental methods for DTI identification are extremely costly and time-consuming, computational DTI prediction methods have received a growing popularity in literature. The majority of existing computational methods treat DTI prediction as a binary classification task, where known DTIs are labeled as positive and unknown DTIs are labeled as negative [1]. To address the imbalanced problem arised from the binary classification scheme, many research has attempted to extract a subset of reliable negative samples, e.g. by random sampling [4] or by **P**ositive **U**nlabel **L**earning (PU Learning) [2].

Instead of labeling the unknown DTIs as negative, we argue that it is more natural to consider the unknown DTIs, i.e. DTIs that are neither identified in vivo to be positive nor experimentally validated to be negative (non-interacting drug-target pairs), as missing labels. Furthermore, our assumption in this work is that labels are not missing at random. This is an intuitive and reasonable assumption, because researchers will use their domain expertise to filter DTIs with a high possibility to be positive and prioritize validations for these DTIs in vivo. For example, researchers find the efficacy target of a drug based on principles of biochemistry, biophysics, genetics and chemical biology. If ample evidences exist to support positive interactions with the target, then the possibility of a positive DTI is high, and the researchers are likely to conduct in vivo experiments. On the contrary, drug target interactions that are less likely to be positive are more likely to be ignored by researchers and their labels are likely to be missing.

1.1 Contribution

Our chief contribution in this work is a novel **F**actorization with **N**on-random **M**issing **L**abels model (FNML). To the best of our knowledge, this is the first time missing not at random theory is applied in DTI identification. The inputs of FNML are feature vectors of drugs and targets, the partially observed labels, and the fully observed responses (i.e. labels are given or missing). The feature vectors are learnt and integrated from heterogenous sources. The labels and responses are binary variables. The FNML model mimics the probabilistic procedures to generate labels from feature vectors and responses from labels. Specifically, the labels are related to feature vectors of both drugs and targets, and a hidden matrix mapping from the drug features to target features. The possibility of giving a response is associated with the sign of the label.

We conduct experiments on the latest DTI database. **A**rea **U**nder **R**eceiver **O**perating **C**haracteristic curve (AUROC) and **A**rea **U**nder **P**recision **R**ecall curve (AUPR) are the most commonly adopted metrics to evaluate DTI prediction performance. Our model achieves best AUPR and AUROC results on both full dataset and sample sets with balanced numbers of positive and negative labels.

1.2 Related Work

One component of our work (i.e labels are generated by feature vectors learnt and fused from heterogenous information networks) is inspired by a recent work DTINet [4]. However, there are three key differences between our work and DTINet. (1) DTINet is based on deterministic matrix factorization, our work is based on probabilistic factor models. For example, the hidden feature space mapping matrix, labels, and responses are all random variables. This setting enables the FNML model to regulate the parameters (i.e. hidden feature space mapping matrix) by introducing appropriate priors. Therefore, performance on sparse dataset is improved. (2) DTINet is based on randomly missing responses, i.e. it samples uniformly a set of unknown DTIs as negative sample, while FNML is based on missing not at random theories. Statistical theory in [3] shows that applying a model based on missing at random assumptions can lead to biased parameter estimation on data sets with missing not at random entries. (3) DTINet adopts only a subset of unknown DTIs to preserve a balanced number of positive and negative samples, while our model uses all information in the data set.

We also want to distinguish our work with another line of research. Usually only positive DTIs are deposited in known databases. Due to the lack of negative samples, PU learning has been employed in DTI identification, e.g. to facilitate negative sample extraction [2].

Conference'17, Washington, DC, USA
2016. 978-x-xxxx-xxxx-x/YY/MM…$15.00
DOI: 10.1145/nnnnnnn.nnnnnnn

PU learning does not explicitly associate the status of an instance (i.e. being labeled or unlabeled) with the value of its label. We also want to mention here that, although we experiment with datasets where only positive DTIs are deposited, FNML is extendable without difficulty to databases where positive and negative DTIs are available. Thus our model is applicable in more scenarios.

2 THE PROPOSED METHOD

We start with the problem definitions and notations in Sec. 2.1. We then describe the proposed model FNML in Sec. 2.2. Finally we present the inference algorithm in Sec. 2.3.

2.1 Preliminaries

DTI identification is often modeled as a binary classification task. Formally, we are given $P \in \mathcal{R}^{N \times M}$ a set of DTI labels, where $p_{i,j} = 0$ indicates a negative interaction between drug i and target j, $p_{i,j} = 1$ indicates a positive DTI, the feature vectors on drug side $X \in \mathcal{R}^{N \times K}$, where $x_{i,k}$ represents drug i's weight on drug feature k, the feature vectors on target side $Y \in \mathcal{R}^{M \times L}$, where $y_{j,l}$ represents target j's weight on target feature l. The problem is to predict for a new drug-target pair $< i', j' >$, the possibility of a positive DTI $p(p_{i',j'} = 1)$.

It is worthy to note that many previous studies have shown that extracting features X, Y from heterogenous data sources are beneficial. Similar to DTINet [4], we use a compact feature expression learnt from aggregating diffusion probabilities on multiple information networks. For example, on the drug side, we can collect drug side-effect network, drug similarity network and so on. On the target side, we can collect protein disease network, protein protein network and so on. The feature learning phase is beyond the scope of this paper. We refer the readers to [4].

In addition to the features X, Y and labels P, we make one essential modification to the problem definition. We assume that the inputs also contain responses $R \in \mathcal{R}^{N \times M}$, where $R_{i,j} = 0$ indicates an unknown DTI, $R_{i,j} = 1$ indicates a verified DTI (positive or negative). For positive responses $R_{i,j} = 1$, the labels $P_{i,j}$ are observed. For negative responses $R_{i,j} = 0$, the labels are hidden and unknown.

2.2 FNML Model

We use a factor model, depicted in Fig. 1. The features X, Y are in different dimensions. To associate the drug features with the target features, we introduce a hidden matrix $Z \in \mathcal{R}^{K \times L}$, where $Z_{k,l}$ is a projection that maps the drug feature k to the target feature l. We assume that Z is sampled from a Gaussian distribution,

$$\forall k, l, Z_{k,l} \sim \mathcal{N}(0, \sigma^2), \tag{1}$$

where σ^2 is the variance. We use zero mean to favor sparse feature mapping, i.e. a drug feature k is associated with a few target features.

We then assume that the binary label $P_{i,j}$ is generated from the following process:

$$\forall i, j, p(P_{i,j} = 1|X, Y, Z) = \frac{1}{1 + \exp{(-XZY)_{i,j}}}. \tag{2}$$

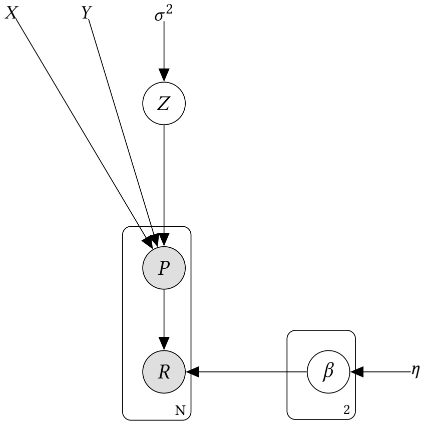


Figure 1: Graphical Representation of the FNML model

The binary response is sampled from a Bernoulli distribution. The parameters of the Bernoulli distribution are related to the value of each $P_{i,j}$. Therefore we define $\beta_p \in \mathcal{R}^2, p \in \{0, 1\}, \forall p, \beta_{p,0} > 0, \beta_{p,1} > 0, \beta_{p,0} + \beta_{p,1} = 1$, we have:

$$\forall p \in \{0, 1\}, \beta_p \sim Beta(\eta), \tag{3}$$

$$\forall i, j, R_{i,j} \sim Bern(\beta_{P_{i,j},1}), \tag{4}$$

where η is the hyperparameter for the Beta distribution.

2.3 Inference

The objective is to maximize the likelihood which consists of two terms. The first term is on partial observations, i.e. $R_{i,j} = 0$ and $P_{i,j}$ unknown. The second term is on full observations, i.e. $R_{i,j} = 1$ and known $P_{i,j}$.

$$\mathcal{L} = \sum_{R_{i,j}=0} \log p(R_{i,j}|X, Y, \sigma^2, \eta) + \sum_{R_{i,j}=1} \log p(R_{i,j}, P_{i,j}|X, Y, \sigma^2, \eta) \tag{5}$$

Direct optimization for both terms in Equ. 5 is intractable, as they involve integration over continuous hidden variables. For example, $p(R|X, Y, \sigma^2, \eta) = \int_{P,Z,\beta} p(R|P, \beta, \eta)p(P|X, Y, Z)p(Z|\sigma^2)p(\beta|\eta)$. We employ variational inference to infer the parameters. That is, we use the mean field assumption to factorize the posterior distribution:

$$q(P, Z, \beta|R, X, Y, \sigma^2, \eta) = q(P|\theta)q(Z|\mu, \upsilon)q(\beta|\rho), \tag{6}$$

It is convenient if $q(P|\theta), q(Z|\mu, \upsilon), q(\beta|\rho)$ are exponential distributions. We approximate the sigmoid function in Equ. 2 by an exponential distribution. We use the property that any sigmoid function $\sigma(\cdot)$ has a lower bound:

$$q(P|\theta) = \sigma(\theta) \geq \sigma(\zeta) \exp{(\theta - \zeta)/2 - \lambda(\zeta)(\theta^2 - \zeta^2)}, \tag{7}$$

where $\lambda(\zeta) = [\sigma(\zeta) - 1/2]/[2\zeta]$.

As shown in Alg. 1, in each iteration of the inference we alternatively optimize the variational parameters for $q(Z|\mu, \upsilon), q(\beta|\rho), q(P|\theta)$ and the parameters for the lower bound $\sigma(\zeta)$. We divide the data objects into two disjoint sets, $s_1 = \{(i, j) \in \mathcal{R}^{N \times M}|R_{i,j} = 1\}$, and $s_2 = \{(i, j) \in \mathcal{R}^{N \times M}|R_{i,j} = 0\}$. In each iteration, we first obtain the optimal θ, μ, v, ρ and then we update ζ. The iteration is repeated until convergence is achieved.

```
input  : P, R, X, Y
output : $\mu$, $v$, $\rho$, $\theta$, $\zeta$
initialization;
repeat
    for $Z_{k,l} \in Z$ do
        $\mu_{k,l} \leftarrow \frac{\sum_{(i,j)\in s_2}(\theta_{i,j}-\frac{1}{2})X_{i,k}*Y_{j,l}+\sum_{(i,j)\in s_1}\frac{1}{2}X_{i,k}*Y_{j,l}}{2*(\sum_{i,j}\lambda(\zeta_{i,j})X_{i,k}^2 Y_{j,l}^2+\frac{1}{\sigma^2})}$;
        $v_{k,l} \leftarrow \frac{1}{\sqrt{2*(\sum_{i,j}\lambda(\zeta_{i,j})X_{i,k}^2 Y_{j,l}^2+\frac{1}{\sigma^2})}}$;
    end
    for $\beta$ do
        $\rho_{0,0} \leftarrow \eta_{0,0}$;
        $\rho_{0,1} \leftarrow \sum_{(i,j)\in s_2}(1-\theta_{i,j})+\eta_{0,1}$;
        $\rho_{1,0} \leftarrow |s_1|+\eta_{1,0}$;
        $\rho_{1,1} \leftarrow \sum_{(i,j)\in s_2}\theta_{i,j}+\eta_{1,1}$;
    end
    for $(i,j) \in s_2$ do
        $l_1 = exp(\psi(\rho_{1,1})-\psi(\rho_{1,0}+\rho_{1,1})+X_i\mu Y_j^T)$;
        $l_2 = exp(\psi(\rho_{0,1})-\psi(\rho_{0,0}+\rho_{0,1}))$;
        $\theta_{i,j} \leftarrow \frac{l_1}{l_1+l_2}$;
    end
    for $(i,j) \in s_1 + s_2$ do
        $\zeta_{i,j} \leftarrow \left|X_i\mu Y_j^T\right|$;
    end
until convergence;
```

Algorithm 1: Inference for FNML

3 EXPERIMENT

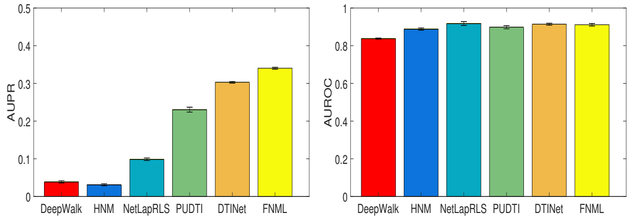
3.1 Experimental Setup

Datasets. We use the same datasets as in [4]: i.e. the drug-target interaction labels are obtained from the latest version of DrugBank (version 3.0); the feature vectors X are extracted from four heterogenous networks: the drug-drug interaction network (dd), the drug-disease network (di), the drug-side-effect network (de) and the drug structure similarity network (ds); the feature vectors Y are extracted from three heterogenous networks: the protein-disease association network (pd), the protein-protein network (pp) and the protein sequence similarity network (ps). As shown in Tab. 1, in the full data set, only 0.18% of the drug-target interactions are labelled as positive, none is labelled as negative. As in [4], we also construct a sample dataset, where all the positive interactions are reserved and an equal number of unknown interactions are sampled to be negative.

Table 1: Statistics of the datasets

Data	#Drugs	#Targets	#Positive	#Negative	#Unknown
Full	708	1,512	1,923	0	1,068,573
Sample	708	1,512	1,923	1,923	1,066,650

Evaluation. Throughout the experiment section, the major evaluation metric is **A**rea **U**nder **P**recision **R**ecall curve (AUPR), which is commonly adopted in bioinformatic studies. An auxiliary evaluation metric is **A**rea **U**nder **ROC** curve (AUROC).



(a) AUPR (b) AUROC

Figure 2: On the full dataset, FNML significantly boosts AUPR while obtaining comparable AUROC.

3.2 Results and Analysis

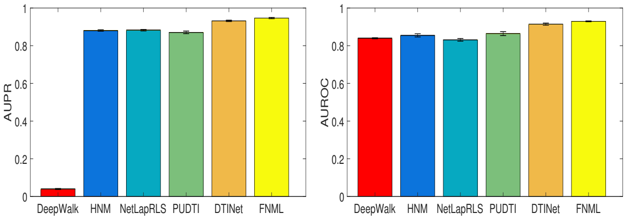
DTI Prediction Performance. We first evaluate the accuracy of DTI prediction of the proposed FNML model. The number of dimensions for drug features are $K = 100$, for target features $L = 400$, hyper-parameters are $\sigma^2 = 1, \eta_{0,0} = 1, \eta_{0,1} = 1, \eta_{1,0} = 1, \eta_{1,1} = 1$.

We compare our FNML model with 5 state-of-the-art methods: (1) DeepWalk [7]: a similarity-based drug-target prediction method that enhances similarity computation by deep learning method within a linked tripartite network. (2) HNM [6]: a network model in which strength between a disease-drug pair is calculated through an iterative algorithm on the heterogeneous graph that also incorporates drug-target information. (3) NetLapRLS [5]: a manifold regularization semi-supervised learning method. (4) PUDTI [2]: an SVM-based optimization model that is trained on negative samples extracted based on positive-unlabeled learning. (5) DTINet [4]: a regression model that learns feature space mapping Z by the loss function $\min_Z \sum_{i,j}(P_{i,j}-(XZY)_{i,j})^2$. We do not change the default settings for all the above comparative methods.

We perform the evaluation on two datasets. The first one is on the full dataset, i.e. we randomly segment the whole data set to 10 divisions and conduct 10-fold cross-validation. The second one is on the sample dataset, i.e. keeping the ratio of positive and negative samples to $1:1$, we conduct random sampling for 10 times and the reported results are averaged over the 10 sets.

The comparative performance on the full dataset are shown in Fig. 2. We can see that (1) FNML model significantly boosts the AUPR performance by 12.35%, compared with the best of state-of-the-art methods. The best comparative method is DTINet, which achieves a 30.29% AUPR. Our FNML model obtains a 34.03% AUPR. As AUPR is well regarded to be a more robust and accurate evaluation metric than AUROC [4], this observation demonstrates the potential of our model. (2) Most of the state-of-the-art methods yield very low AUPR results on the full dataset. This observation again reveals that obtaining a high AUPR performance is challenging on the full dataset. (3) In term of AUROC, the best result is obtained by NetLapRLS. However, the best comparative result is 91.78%, while FNML produces a comparable 91.12% AUROC.

The comparative performance on the sample dataset are shown in Fig. 3. We can see that (1) FNML model achieves better AUPR than all state-of-the-art methods. The best comparative method is DTINet, which achieves 93.20%. Our FNML model obtains a



(a) AUPR (b) AUROC

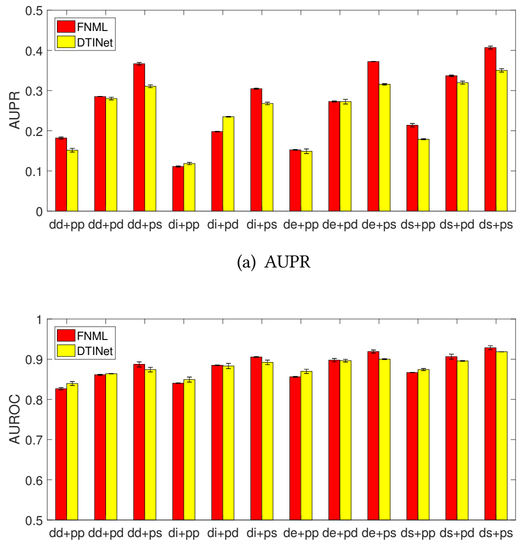
Figure 3: On the sample dataset, FNML outperforms state-of-the-art methods in terms of both AUPR and AUROC.

94.66% AUPR. (2) Most of the state-of-the-art methods have a higher AUPR result on the sample dataset than the full dataset, due to the balanced ratio of positive and negative samples. (3) FNML model outperforms all state-of-the-art models in AUROC performance. The best comparative method is again DTINet, which achieves 91.41%. Our FNML model obtains a 92.93% AUROC. (4) Surprisingly, DeepWalk has a very low AUPR performance on the sample set. A possible reason is that the network representation extracted by deepwalk is based on homogeneous network structure, and thus is not accurate. Moreover, we use the public open source codes provided by authors in [7], which are not specifically tunned for the DrugBank dataset.

Stability. We next study the stability of our FNML model. We use various combination of X and Y as inputs. That is, we extract X from the four networks on the drug side (i.e. dd,di,de,ds) respectively, extract Y from the three networks on the protein side (i.e. pp, pd, ps) respectively, and use the 12 combinations as inputs to train the model. The predictions are tested on the full dataset.

We compare the AUPR and AUROC performance of FNML and DTINet. As shown in Fig. 4, FNML outperforms DTINet in most cases. FNML generates better AUPR results for 10 feature combinations out of 12. In term of AUROC, FMNR is better for 7 feature combinations. The result shows that the performance improvement is stable. Change of feature representations does not affect FNML’s ability to learn a better feature mapping space.

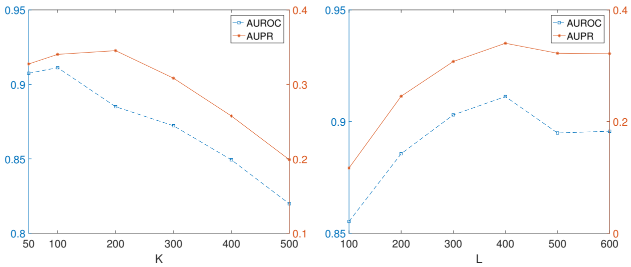
Number of dimensions. Finally we study the effects of number of dimensions K, L. We first fix $L = 400$ and tune from $K = 50$ to $K = 500$.We can see from Fig. 5(a) that the best number of drug features is around 100. Then, we fix $K = 100$ and tune from $L = 100$ to $L = 600$. As shown in Fig. 5(b), the best number of target features is 400. An appropriate number of drug features is important. When the number of drug features is too large, i.e. $K > 200$, we observe a descent fall in both AUPR and AUROC. However, the model performance is less sensitive to the number of target features. For $L > 500$, AUPR and AUROC remain the same.



(a) AUPR

(b) AUROC

Figure 4: Our model is stable with different feature inputs.



(a) K number of drug features (b) L number of protein features

Figure 5: AUPR and AUROC performance of our model with different number of drug and protein features.

4 CONCLUSION

We propose a novel DTI prediction model based on the assumption that unknown DTI labels are missing not at random. By associating the status of a DTI being labelled or unknown to the sign of the DTI label, our proposed FNML model can learn a better feature mapping from drug feature space to target feature space. We experimentally demonstrate that FNML outperforms state-of-the-art computational DTI identification methods. This work sheds some insights into fully exploiting the information in unknown DTIs. Our further directions include improving the factorization framework and analyzing the missing mechanisms.

REFERENCES

[1] Ding, H., Takigawa, I., Mamitsuka, H., and Zhu, S. Similarity-based machine learning methods for predicting drug-target interactions: a brief review. In *Briefings in bioinformatics*, 15(5), 734-747.

[2] Peng L, Zhu W, Liao B, et al. Screening drug-target interactions with positive-unlabeled learning. In *Scientific Reports*, 2017, 7(1): 8087.

[3] R. J. A. Little and D. B. Rubin. Statistical Analysis with Missing Data, 1987.

[4] Luo Y, Zhao X, Zhou J, et al. A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. In *Nature Communications*, 2017, 8(1).

[5] Xia Z, Wu L Y, Zhou X, et al. Semi-supervised drug-protein interaction prediction from heterogeneous biological spaces. In *Bmc Systems Biology*, 2010, 4(S2):1-16.

[6] Wang W, Yang S, Zhang X, et al. Drug repositioning by integrating target information through a heterogeneous network model. In *Bioinformatics*, 2014, 30(20):2923-2930.

[7] Zong N, Kim H, Ngo V, et al. Deep Mining Heterogeneous Networks of Biomedical Linked Data to Predict Novel Drug-Target Associations. In *Bioinformatics*, 2017, 33(15).