Drug Target Interaction Prediction based on Missing not at Random Labels

Anonymous Author(s)

ABSTRACT

We introduce a novel probabilistic model, **F**actorization with **M**issing **N**ot at **R**andom **L**abels (FMNRL), for **D**rug-**T**arget **I**nteraction (DTI) prediction. Unlike previous studies which label unknown DTIs as negative samples, we treat the unnown DTIs as missing not at random responses. FMNRL assumes the labels are probabilistically generated from feature vectors of both drugs and targets, and a hidden matrix mapping from the drug features to target features. By associating the possibility of missing response to the possibility of a negative label, FMNRL can better learn the hidden feature space mapping and thus provide more accurate DTI predictions. Experimental results show that FMNRL achieve significant improvement in term of AUROC and AUPR.

ACM Reference Format:
Anonymous Author(s). 2018. Drug Target Interaction Prediction based on Missing not at Random Labels. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 2 pages. https://doi.org/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 mojority 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 []. To address the imbalanced problem arised from the binary classification scheme, many research works attempt to extract a subset of reliable negative samples, e.g. by random sampling [3] or by **U**nlabel **L**earning (PU Learning) [1].

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 responses (i.e. the labels are missing). 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 possiblity to be positive and priorize validations for these DTIs in vivo. For example, if researchers prefer to validate drugs with certain pharmacokinetic interactions, i.e. one drug affects the in vivo absorption, distribution, metabolism, or excretion of the target, then the unknown DTIs are not missing at random because drugs without pharmacokinetic interactions with the target are less likely to be positive and hence more likely to be missing.

Conference'17, July 2017, Washington, DC, USA
2018. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1.1 Contribution

Our chief contribution in this work is a novel **F**actorization with **M**issing **N**ot at **R**andom **L**abels model (FMNRL). To the best of our knowledge, this is the first time missing not at random theory is applied in DTI identification. The inputs of FMNRL are feature vectors of drugs and targets which are leant and integrated from heterogenous data sources, the partially observed labels (i.e. positive or negative), and the fully observed reponses (i.e. given or missing). The FMNRL model mimics the probabilitic procedures to generate the labels from feature vectors and the 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 to the possibility of a positive label.

We conduct experiments on a large-scale DPI database. Our model achieves significantly better AUPR result and comparable AUROC result to the best of related works [3]. In the biomedical field, AUROC is considered to be a more robust and better assessment than AUROC [3]. Thus our improvment in AUPR is promising.

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 [3]. However, there are three key differences between our work and DTINet. (1) DTINet is based on deteministic 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 FMNRL model to regulate the parameters (i.e. hidden feature space mapping matrix) by introducing approapriate priors and improves performance on sparse data set. (2) DTINet is based on randomly missing responses, i.e. it samples uniformly a set of unknown DTIs as negative sample, while FMNRL is based on missing not at random theories. Statistical theory in [2] 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, recently **P**ositive **U**nlabel **L**earning (PU Learning) is empolyed in DTI identification, e.g. to facilitate negative sample extraction [1]. PU learning does not explicitly associate the status of an instance (i.e. being positive or unlabel) with the value of its hidden label. We also want to mention here that, although we experiment with datasets where only positive DTIs are deposited, FMNRL is extendable without difficulty to databases where positive and negative DTIs are available. Thus our model is applicable in more scenarios.

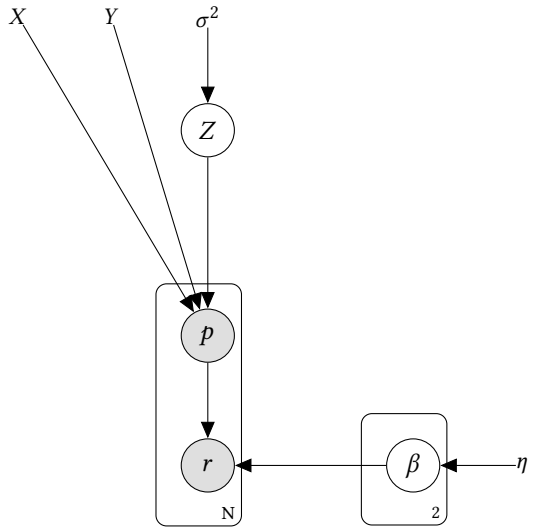

Figure 1: Graphical Representation of the FMNRL model

2 THE PROPOSED METHOD

2.1 Preliminaries

Given a set of DTI labels $P = \{p\}$, where $p \in \{0, 1\}$. responses $P = \{p\}$, where $p \in \{0, 1\}$

2.2 Model

$$z \sim \mathcal{N}((0, \sigma^2) \tag{1}$$

$$p(p = 1) = \frac{1}{1 + \exp(xzy)} \tag{2}$$

$$\beta \sim Beta(\eta) \tag{3}$$

$$r \sim Bern(\beta_p) \tag{4}$$

2.3 Inference

3 EXPERIMENT

3.1 Experimental Setup

3.2 Results and Analysis

4 CONCLUSION

REFERENCES

[1] Peng L, Zhu W, Liao B, et al. Screening drug-target interactions with positive-unlabeled learning. In *Scientific Reports*, 2017, 7(1): 8087.
[2] R. J. A. Little and D. B. Rubin. Statistical Analysis with Missing Data, 1987.
[3] 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).