Neurocomputing ∎ (∎∎∎∎) ∎∎∎–∎∎∎

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Predicting potential side effects of drugs by recommender methods and ensemble learning

Wen Zhang [a,b,*], Hua Zou [a], Longqiang Luo [c], Qianchao Liu [a], Weijian Wu [a], Wenyi Xiao [a]

[a] School of Computer, Wuhan University, Wuhan 430072, China
[b] Research Institute of Shenzhen, Wuhan University, Shenzhen 518057, China
[c] School of Mathematics and Statistics, Wuhan University, Wuhan 430072, China

ARTICLE INFO

Article history:
Received 18 January 2015
Received in revised form
29 June 2015
Accepted 23 August 2015

Keywords:
Drug side effects
Recommender system
Ensemble learning
Restricted Boltzmann machine

ABSTRACT

Drugs provide help and promise for human health, but they usually come with side effects. Predicting side effects of drugs is a critical issue for the drug discovery. Although several machine-learning methods were proposed to predict the drug side effects, it remains the space for the improvement. To the best of our knowledge, many side effects are not detectable in clinical trials until drugs are approved, thus predicting potential or missing side effects based on the known side effects is important for the post-marketing surveillance. In order to solve this specific problem, we formulate approved drugs, side effect terms and drug–side effect associations as a recommender system, and transform the problem of predicting side effects into a recommender task. Two recommender methods, i.e. the integrated neighborhood-based method and the restricted Boltzmann machine-based method, are designed to make predictions. Further, in order to achieve better performances, we combine proposed methods and existing methods of the same type to develop ensemble models. Compared with benchmark methods, the proposed methods and the ensemble method lead to better performances, and the statistical analysis demonstrates the improvements are significant (p-value < 0.05). In conclusion, the integrated neighborhood-based method, the restricted Boltzmann machine-based method and the ensemble method are promising tools for the side effect prediction. The source codes and datasets are provided as the supplementary.

1. Introduction

A drug is a chemical substance for the treatment, cure, prevention, diagnosis of diseases, etc. Despite providing promise for the human health, drugs usually come with side effects or adverse reactions. Considering the fact that side effects may lead to serious consequences, such as failure in drug development or drug withdrawal, side effects have become the public health concern, and identifying side effects is an important topic in the drug discovery.

In wet experiments, researchers conduct the counter screen of compounds against a large number of enzymes and receptors in vitro to identify drug side effects. However, wet experiments are generally time-consuming and laborious. With the increase of drug data, computational methods were proposed to reduce the cost and speed up drug development. Early computational methods utilized the structure–activity relationship or quantitative structure–property relationship for prediction. For example, Fliri [1] transformed the side effect data from prescription drug labels into effect spectra, and then diagnosed drug side effects. Fukuzaki [2] predicted side effects based on sub-pathways that share correlated modifications of gene-expression profiles. Hammann [3] presented a structure–activity relationship analysis of side effects in the central nervous system, liver and kidney. Scheiber [4] took analysis to identify chemical substructures associated with side effects. In general, these methods focus on the case study.

In recent years, machine-learning methods were applied to the side effect prediction, extending works from case study to more data. Generally speaking, these methods take into account the chemical and biological features of drugs (i.e. chemical structures, target protein, protein–protein interaction, gene ontology annotations, etc.), and machine learning techniques are adopted to build the relationship between drug-related features and drug side effects. Huang [5] used the drug targets, protein–protein

* Corresponding author at: School of Computer, Wuhan University, Wuhan 430072, China.
E-mail addresses: zhangwen@whu.edu.cn (W. Zhang), zouhua08@gmail.com (H. Zou), lqluo@whu.edu.cn (L. Luo), Ritty@whu.edu.cn (Q. Liu), wuweijian@whu.edu.cn (W. Wu), icy@whu.edu.cn (W. Xiao).

http://dx.doi.org/10.1016/j.neucom.2015.08.054
0925-2312/

interactions and gene ontology annotations as features, and adopted the support vector machine (SVM) and the logistic regression to construct predictors. Pauwels [6] utilized chemical structures, and four machine-learning methods (k-nearest neighbor, support vector machine, ordinary canonical correlation analysis and sparse canonical correlation analysis) were adopted to train prediction models. Mizutani and Yamanishi [7,8] combined chemical structures and target proteins, and adopted the sparse canonical correlation to build prediction models. Liu [9] integrated a wide variety of drug-related features, and then built the prediction models by respectively using five machine-learning classifiers (logistic regression, naive Bayes, k-nearest neighbor, random forest and SVM). Bresso [10] made use of decision trees and inductive logic programming to identify and characterize side-effect profiles shared by several drugs. Huang [11] combined chemical structures and protein–protein interactions, and then built SVM-based predictors. Jahid [12] used chemical structures as features, and combined the results from different binary classification models for prediction. Liu [13] determined molecular predictors of adverse drug reactions with causality analysis.

In addition to methods based on the drug-related chemical and biological features, researchers put efforts to predict potential side effects based on known side effects. In Liu's work [9], known side effects were encoded as feature profiles, and experimental results demonstrated that this feature even yielded much better results than chemical and biological features. Cheng [14] represented drugs, side effect terms and known side effects as a bipartite. In Cheng's work, known side effects served as the initial resources, and the resource allocation-based method named 'network inference method (NIB)' was adopted to infer potential side effects from known side effects. There are some reasons for developing the known side effect-based prediction. Firstly, many side effects are not detected in clinical trials until drugs are approved, and predicting potential or missing side effects is critical for post-marketing surveillance. Secondly, most existing methods utilize drug-related biological and chemical features to make predictions, but the information about features are not always available. The state-of-the-art methods cannot work without the necessary information.

Inspired by the pioneering works, we attempt to predict the missing or potential side effects of approved drugs by using known side effects. In this paper, we formulate approved drugs, side effect terms and drug–side effect associations as a recommender system, thus then transform the potential side effect prediction into a recommender task. In this way, we attempt to solve the original problem in the frame of recommender system. In order to make predictions, we design two recommender methods, namely the integrated neighborhood-based method (INBM) and the restricted Boltzmann machine-based method (RBMBM). INBM is an extension of the classic neighborhood-based recommender method, and it utilizes similar drugs' known side effects for prediction; RBMBM constructs a two-layer network, which learns the probability distribution governing drug–side effect associations. To achieve better performances, proposed methods and existing methods of the same type are integrated to develop ensemble models. Compared with the state-of-the-art methods, INBM, RBMBM and the ensemble method yield better AUPR scores and AUC scores on the benchmark datasets, and the statistical analysis demonstrates that the improvements are significant (p-value < 0.05). The source codes and datasets are provided as the supplementary.

2. Materials and methods

2.1. Datasets

There are several databases for drugs, side effects and drug-related information, namely SIDER [15], PubChem Compound [16,17], DrugBank [18–21] and KEGG DRUG [22]. SIDER database contains marketed medicines and their adverse drug reactions. In addition, important information about the medicines, i.e. side effect frequency, side effect classifications and drug–target relations, are provided. PubChem Compound database contains validated chemical depiction information that describes substances in PubChem Substance. DrugBank database is a bioinformatics and cheminformatics resource that includes detailed drug (i.e. chemical, pharmacological and pharmaceutical) data with comprehensive drug target (i.e. sequence, structure, and pathway) information. KEGG DRUG is a comprehensive drug information resource for approved drugs in Japan, USA, and Europe unified based on the chemical structures and/or the chemical components, and associated with target, metabolizing enzyme, and other molecular interaction network information.

To date, several datasets, i.e. Pauwels's dataset [6], Mizutani's dataset [7] and Liu's dataset [9], were compiled from above databases, and were used to develop the state-of-the-art methods. More importantly, these datasets are publicly available. Hence, we adopt these datasets as the benchmark datasets to evaluate and compare methods. Table 1 displays the details of three datasets, covering the number of approved drugs, number of side effect terms, and the number of side effects of the approved drugs (also known as 'drug–side effect associations'). These datasets also contain biological or chemical information about drugs. Since our work is to infer the potential side effects from known side effects, we only utilize known side effects of approved drugs to build prediction models.

2.2. Problem description

For a dataset with n drugs and m side effect terms, the drug set and side effect set are respectively denoted as $\boldsymbol{D} = \{\boldsymbol{d}_1, \boldsymbol{d}_2, \ldots \boldsymbol{d}_n\}$ and $\boldsymbol{S} = \{\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_m\}$. The side effects of approved drugs, also known as drug–side effect associations, are naturally represented as an $n \times m$ adjacent matrix. As shown in Table 2, the binary value '0' or '1' for the entry $M_{i,j}$ represents the absence or presence of side effect s_j for drug d_i.

We design a recommender system, by incorporating drugs, side effects and drug–side effect associations. As shown in Fig. 1, the recommender system is illustrated by a bipartite network, in which drugs and side effects are represented as nodes of two layers. Drug nodes are linked to side effect nodes according to the known drug–side effect associations. Therefore, inferring potential side effects from known side effects is equivalent to a

Table 1
The details of benchmark datasets.

Datasets	# drugs	# side effect terms	# drug–side effect associations	Additional information
Pauwels's dataset	888	1385	61,102	Substructures
Mizutani's dataset	658	1339	49,051	Substructures, target proteins
Liu's dataset	832	1385	59,205	Substructures, targets, transporters, enzymes, pathways, etc.

Please cite this article as: W. Zhang, et al., Predicting potential side effects of drugs by recommender methods and ensemble learning, Neurocomputing (2015), http://dx.doi.org/10.1016/j.neucom.2015.08.054