ATLANTIS PRESS

Advances in Engineering Research, volume 113

International Forum on Mechanical, Control and Automation (IFMCA 2016)

Drug Side-effect Prediction based on Comprehensive Drug Similarity

Chengcheng Sun[1,a], Yi Zheng[2,b], Yan Jia[3,c] and Liang Gan[4,d]

[1,2,3,4] *School of Computer Science*, National University of Defense Technology

Changsha, Hunan Province, China

[a]sunchengcheng12@163.com, [b]justice131@163.com

Keywords: drug side-effect prediction; comprehensive drug similarity; drug chemical structure; drug target protein.

Abstract. Drug side-effects have gained attention from the whole society because of significant morbidities and mortalities they caused. Therefore, the early detection of side-effects, before reaching the clinical stages, is important yet challenging. In this paper, we proposed a comprehensive similarity framework to measure the similarity between drugs and developed a side-effect prediction approach based on it. The new comprehensive similarity integrates drug formula similarity, drug chemical structure similarity, drug substituent similarity, drug target similarity and drug side-effect similarity in a unified framework to capture drug features from different perspectives. Extensive comparison experiments on real drug side-effect prediction demonstrate that our proposed approach is competent to drug side-effect prediction and outperforms other state-of-the-art approaches.

Introduction

Drug side-effects or adverse drug reactions, are phenotypic responses of the human organism to drug treatments [1]. They have gained the public attention due to the significant morbidities and mortalities they caused. It is estimated that serious drug side-effects are responsible for 100,000 deaths each year and rank the fourth among the leading causes of deaths across the USA [2]. Approximately 40,000 visits to general practitioners and 190,000 hospital admissions every year are caused by side-effects in Australia with only a small population of about 23 million. As well, lots of money has been put on side-effects. Taking Australia as an example, around 660 million dollars were spent on medicine-related hospitalizations including side-effects and medication errors in 2009 [3]. Thus, to identify the potential side-effects before their appearance on the market and further leading to serious loss is of great importance. Though they are experimental approaches which attempts to anticipate side-effects by testing compounds with in vitro biochemical and cellular assays (i.e. preclinical safety profiling), however, experimental side-effect detection remains challenging due to its high cost and low efficiency. Recently, a few computational methods have been proposed for drug side-effect predictions. These methods can be classified as target protein-based methods, pathway-based methods and chemical structure-based methods.

The principle of target protein-based methods is to relate drug side-effects to its target proteins. In 2008, Monica et al. proposed a method to identify drug targets based on side-effect similarity and verified their findings via vitro binding assays [4]. Yoshihiro and his team built a drug-target interaction search engine by fusion drug side-effects, chemical structures and protein domains [5].

The two findings both leverage side-effects to predict drug-target interactions, demonstrating great prospects to predict drug side-effects via target proteins. Consequently, Yoshihiro's team explored to predict side-effects by combining target proteins and drug structures [6]. Their experimental results show that the prediction accuracy improves owing to integration of target protein information. A triangle-linking method to link drug molecular structure to side-effects via protein targets was proposed by Bender [7]. Through two steps of linking, they achieved ideal results.

The pathway-based methods try to associate drug side-effects with pathways, which involve proteins targeted by the drug. An analysis method named sparse canonical correlation was utilized to extract correlated sets between side effects and target proteins by Sayaka [8]. Their enrichment analyses using KEGG and Gene Ontology showed most of the correlated sets are proteins involved in the same biological pathways. An efficient algorithm called CoopeRative Pathway Enumerator was developed to identify cooperative pathways [9]. Then these cooperative pathways which share common active conditions are used for side-effect prediction. However, the pathway-based methods have limited applicability cause of their heavy dependences on the availability of pathway data.

As the name implies, the chemical structure-based methods relate drug side-effects to drug chemical structures. Josef et al. presented a global linkage analysis to map chemical features to side-effects on a large scale [10]. Their goal simply lies in relating chemical structures to side-effects directly, not intending to understand the mechanistic cause. Based on ordinary canonical correlation analysis method, an improved method i.e. sparse canonical correlation analysis (SCCA) was developed for side-effect profile prediction. 1385 side-effects and 881 chemical substructures are used for drug representation and further relation analysis [11]. Yoshihiro group integrated drug chemical structures and drug target proteins in one framework for side-effect prediction [6]. Extensive comparison experiments showed that their unified framework performed the best.

Most of the state-of-art side-effect prediction approaches focus on mining information from one single source only. However, it is more logical to combine different information sources to avoid bias.

In this paper, we propose a method to predict potential drug side-effects using comprehensive drug similarities and known drug side-effects. Drug formula similarity, drug chemical structure similarity, drug substituent similarity, drug target similarity and drug side-effect similarity all are fused as our comprehensive drug similarities. These different similarities capture features from different perspectives, eliminating bias as possible. The originality of our work lies in the integration of the above five types of similarity into one unified framework. We verify the usefulness of our new method on simulative prediction of 300 side-effects for 1135 drugs from DrugBank, a well-known drug database. Extensive experiments show that the prediction performance improves owing to integration of different similarity information.

This paper is organized as follows. Section II introduces the side-effect prediction method based on our proposed comprehensive drug similarities. Section III details results of our demonstration experiments. Finally, conclusions are drawn in Section IV.

Methods and Materials

A. Data Set

The drug information used in this study was collected from DrugBank [12], a comprehensive drug database. The side effects were crawled from DrugCom [13] and SIDER (version 2) [14]. In this study, we focus on side effects which are related to drugs in the DrugBank database. There are

some side-effects that are related to very few drugs (e.g., excessive thirst, eosinopenia, reticuloendotheliosis and retinal atrophy). Limited information can be provided for such side-effects. Consequently, we removed side-effects which were associated with less than 100 drugs. Finally, we got a dataset consisting of 300 side-effects terms, 1134 drugs and 75,578 associations between drugs and side-effects. The histogram and index-plot of associated drugs for each side-effect are illustrated in the left panel and the right panel of Figure 1 respectively. On average, each side-effect has 251.9 associated drugs and each drug has 66.6 side-effects.

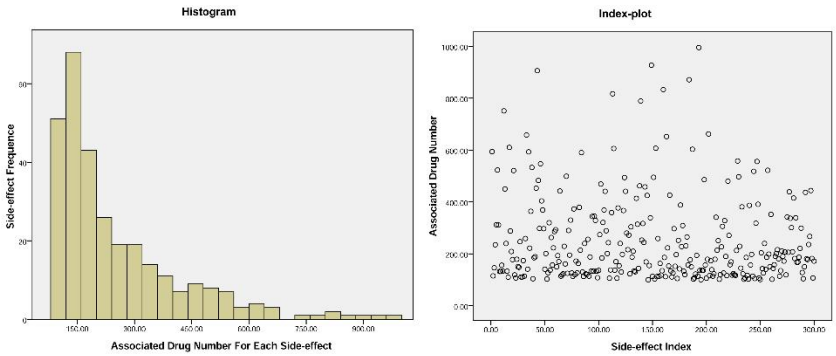

Figure 1. Characteristics of side-effects and associated drugs. The left panel is the histogram of associated drug number for every side-effect and the right panel is the index-plot of the number of associated drug.

B. Comprehensive Similarity

The five types of similarities are integrated as our comprehensive similarity using their mean values. Taking a give drug-drug pair and as an example, the comprehensive similarity is calculated by the following formula:

Where denote drug formula similarity, drug chemical structure similarity, drug substituent similarity, drug target similarity and drug side-effect similarity respectively.

1. Drug Chemical Formula Similarity

The drug chemical formula similarity between drug and is defined as

Where and are the chemical formula element sets of drug and respectively. For instance, for the drug "Abacavir" with formula "$C_{14}H_{18}N_6O$", its chemical formula element set will be {C, H, N, O}. The chemical formula element set represents the chemical elements a drug is composed of. Thus the more elements two drugs share the more similar they are, and vice versa.

2. Drug chemical structure similarity

The chemical structure similarity between drug and is calculated by Tanimoto similarity tool from CDK v1.5.13 [15]. Before calculating, the "simles" string of each drug is parsed and converted to bit fingerprint sets. Then the Tanimoto similarity is calculated as

Where are bit fingerprint sets for drug and respectively; are the bit of ; are bitwise and, or operators; *len* is the length of fingerprint.

3. Drug substituent similarity

The drug substituent similarity is calculated via Jaccard score which is defined as follow:

Where and are the substituent sets of drug and respectively; are union and intersection operators.

4. Drug target similarity

The drug target similarity is defined as:

Where and are the substituent sets of drug and respectively.

5. Drug side-effect similarity

The Drug side-effect similarity is calculated as follow:

Where and are the known side-effect sets of drug and respectively.

C. Prediction Methods

1. Random assignment (Random)

A random assignment classifier is built as our baseline. We randomly assign the label -1/1 to each test drug according to the ratio of -1/1 labels in the training set. For instance, if the ratio for label -1 in training set is 55%, then we assign -1 to 55% of samples in the test set; otherwise 1.

2. Support vector machine (SVM)

SVM is a popular supervised classifier, and it has been widely used in pharmaceutical data analysis [16] and bioinformatics [17] due to its excellent prediction capacity. In this study, in order to predict side-effects of drugs via their similarities, we define their similarity matrix as our self-defined kernel for SVM classification. For a training set of q drugs, the value in the row and column of the similarity matrix is equal to the comprehensive similarity between drug and (1<=i<=q, <=j<=q). Thus, the self-defined kernel is a symmetric matrix with 1 along the main diagonal. For each test drug , it will be represented by a similarity vector , where q is the total number of drugs in training set; is the comprehensive similarity between and the drug in in training set. For each side-effect, one individual SVM classifier has to be built. Thus for our 300 side-effects, we have to build 300 SVM classifier, which will require considerable computational burden.

3. K-Nearest Neighbor (KNN)

KNN is a fundamental and simple classifier, which are commonly first choice while distribution of data is not known [18]. The idea of KNN is to classify an object according to the majority vote of its neighbors. In this study, we use our proposed comprehensive similarity as the distance measure between drugs. Specifically, the distance (i.e. comprehensive similarity) is calculated between the test drug and each drug in the training set. Then the top k drugs, in terms of similarity, are selected as neighbors. Among the k neighbor drugs, if more are associated with one side-effect the test drug will be predicted to cause this side-effect; otherwise not.

Experiments

A. Evaluation Metrics

To evaluate the performance of side-effect prediction, we use prediction, recall and F1-Scoore as our evaluation metrics. Their definitions are as fellows.

Where TP denotes true positives, FP denotes false positive and FN is false negative.

B. Side-effect Prediction Results

1. Parameter Optimization

To achieve better prediction performance of side-effects, parameters of the above three classifiers

are required to be optimized. Specifically, for the random assignment classifier, the ratio of each label is set according to that in the training set; For the SVM classifier, since the prediction task is a multi-class classification problem, thus the SVM type is set as 0. The SVM kernel type is set as 4 (i.e. precomputed kernel) cause the kernel is our comprehensive similarity matrix, and other parameters are default. For KNN classifier, the key parameter is k-value. Thus we carry out extensive experiments to obtain the most proper k-value. Results of the k-value optimization procedure are illustrated in Figure 2. The box-plots of F1-Scores for individual side-effects with different k-values shows that the prediction performance varies a lot with k-value. Note the x-axis denotes the k-value. For example, k50 represent k-value=50. The prediction performance improves slightly with the increase of k-value when k-value is less than 50. However, while k-value is greater than 50, the performance gradually decreases with k-value. Therefore, we choose 50 as our k-value, at which, our KNN classifier achieves the best performance.

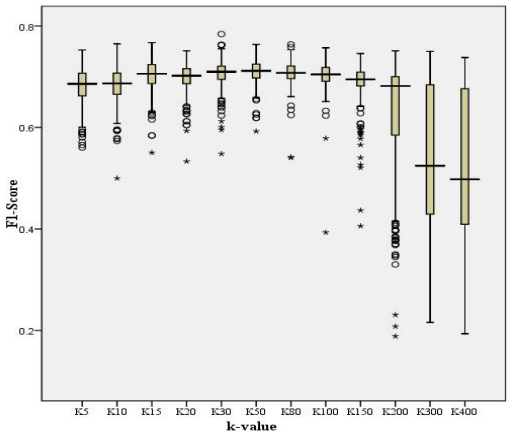

Figure 2. KNN parameter optimization results.

2. Side-effect Prediction

To demonstrate the side-effect prediction performance based on our comprehensive similarity (com), we perform comparison experiments with those chemical similarity (chem), target similarity (tar) and chemical-target (chemTar) similarity based prediction approaches. We test nine approaches: (1) “Random”, (2) “SVM com”, (3) “KNN com”, (4) “SVM chem”, (5) “KNN chem”, (6) “SVM tar”, (7) “KNN tar”, (8) “SVM chemTar” and (9) “KNN chemTar” on their capacities to predict known side-effects. Note that the approaches are named using the combination of similarity type and classifier type used. Taking “SVM com” as an example, it denotes the prediction is performed by SVM based on comprehensive similarity. Com, chem, tar and chemTar are shorted for comprehensive, chemical, target and chemical-target. We conduct fivefold cross-validation as follows: for every side-effect, each drug is flagged as positive if it is associated with the side-effect; otherwise negative. To keep the balance of positive samples and negative samples, we select all drugs from the smaller set between positive drug set and negative drug set, and the same number of drugs from the larger set for further training and test. Then the selected drugs are used as the gold standard set and split into five roughly equal-size subsets. Each subset is used as test set in turn and the remaining four subsets are taken as our training set.

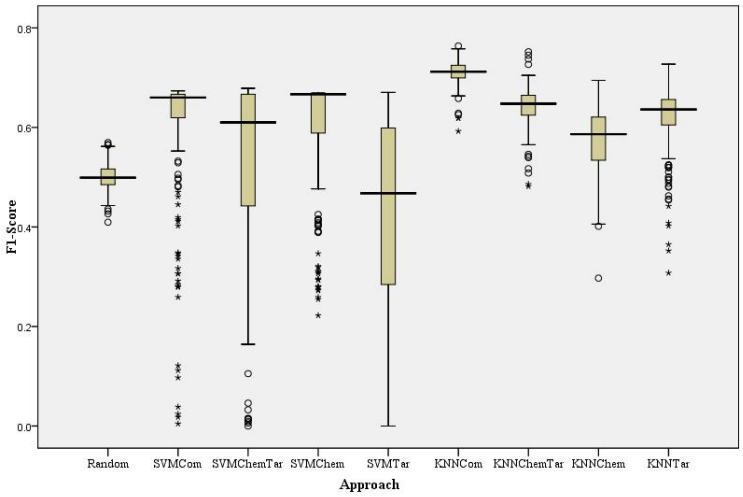

Figure 3. F1-Score box-plots of comparison experiments based on fivefold cross-val

Detailed comparison experiment results are illustrated in Figure 3. From the perspective of the baseline approach "Random", we can see that most of other approaches only except SVMTar all perform better. It indicates that the feasibility of the idea to predict side-effects on drug similarities. The approaches based on our comprehensive similarity, target similarity and chemical-target similarity using KNN classifier achieve better performances than those use SVM classifier (i.e. KNNCom>SVMCom, KNNTar>SVMTar, KNNChemTar>SVMChemTar). However, SVM performs better than KNN in terms of using chemical similarity. This shows that different classifiers are suitable for different distribution of data. Compared with approaches based on chemical similarity, target similarity and chemical-target similarity, our proposed comprehensive similarity based approach performs the best (i.e. SVMCom> SVMChem>SVMChemTar>SVMTar, KNNCom>KNNChemTar>KNNTar>KNNChem). This can be verified by the average precision, recall and F1-Score listed in table 1 as well. From Table 1, we can clearly see that our comprehensive similarity based approach improves by 6.37% (70.91% vs 64.54%) and 1.03% (57.12% vs 56.09%) than the existed second best approach via using KNN and SVM classifiers respectively, in terms of average F1-Score. Analogously, our approach improves by 3.05% (57.57% vs 54.52%) and 0.17% (49.17% vs 46.68%) regarding to average precision, by 28.19% (92.8% vs 79.46%) and 3.23% (79.92% vs 76.69%) regarding to average recall, than the second excellent approach based on KNN and SVM respectively.

Table 1. Performance statistics in fivefold cross-validation

Approach	Average Precision	Average Recall	Average F1-Score
Random	49.92%	49.90%	49.91%
KNNCom	**57.57%**	**92.80%**	**70.91%**
KNNChemTar	**54.52%**	**79.46%**	**64.54%**
KNNChem	53.17%	64.61%	57.82%
KNNTar	53.96%	76.20%	62.23%
SVMCom	**49.17%**	**79.92%**	**57.12%**
SVMChemTar	48.82%	63.62%	53.07%
SVMChem	**49.00%**	**76.69%**	**56.09%**
SVMTar	46.68%	45.45%	42.69%

Conclusion

In this paper, we developed a side-effect prediction approach based on our proposed comprehensive similarities, which integrate drug formula similarity, drug chemical structure similarity, drug substituent similarity, drug target similarity and drug side-effect similarity in a unified framework. These different similarities are capable to capture drug features from different perspectives. Extensive comparison experiments were performed with those approaches based on chemical similarity, target similarity or them together using SVM and KNN classification methods. The experiment results show that our approach is competent in side-effect prediction and achieve the best performance on all the evaluation metrics. In the future, we plan to integrate more similarity sources, e.g. drug therapeutic similarity in our framework to further improve the prediction performance.

Reference

[1] Kuhn M, Campillos M, Letunic I, et al. A side effect resource to capture phenotypic effects of drugs[J]. Molecular Systems Biology, 2010, 6(1):: 343.
[2] Giacomini K M, Krauss R M, Dan M R, et al. When good drugs go bad. Nature[J]. Nature, 2007, 446(7139):975-7.
[3] Semple S J, Roughead E E. Medication safety in acute care in Australia: where are we now? Part 1: a review of the extent and causes of medication problems 2002-2008.[J]. Australia and New Zealand Health Policy, 2009, 6(1):1-12.
[4] Campillos M, Kuhn M, Gavin A C, et al. Drug target identification using side-effect similarity.[J]. Science, 2008, 321(5886):263-6.
[5] Yamanishi Y, Kotera M, Moriya Y, et al. DINIES: drug-target interaction network inference engine based on supervised analysis.[J]. Nucleic Acids Research, 2014, 42(w1):39-45.
[6] Yamanishi Y, Pauwels E, Kotera M. Drug side-effect prediction based on the integration of chemical and biological spaces.[J]. Journal of Chemical Information & Modeling, 2012, 52(12):3284-92.
[7] Bender A, Scheiber J, Glick M, et al. Cover Picture: Analysis of Pharmacology Data and the Prediction of Adverse Drug Reactions and Off-Target Effects from Chemical Structure (ChemMedChem 6/2007)[J]. ChemMedChem, 2007, 2(6):733-733.
[8] Mizutani S, Pauwels E, Stoven V, et al. Relating drug–protein interaction network with drug side effects[J]. Bioinformatics, 2011, 28(18):i522-i528.
[9] Fukuzaki M, Seki M, Kashima H, et al. Side Effect Prediction Using Cooperative Pathways[C]// IEEE International Conference on Bioinformatics and Biomedicine. IEEE, 2009:142-147.
[10] Scheiber J, Jenkins J L, Sukuru S C, et al. Mapping adverse drug reactions in chemical space.[J]. Journal of Medicinal Chemistry, 2009, 52(9):3103-7.
[11] Pauwels E, Stoven V, Yamanishi Y. Predicting drug side-effect profiles: a chemical fragment-based approach[J]. BMC Bioinformatics, 2011, 12(1):1-13.
[12] Law V, Knox C, Djoumbou Y, et al. DrugBank 4.0: shedding new light on drug metabolism[J]. Nucleic Acids Research, 2014, 42(Database issue):1091-7.
[13] Drugs.com. [Online]. Available: http://www.drugs.com
[14] Kuhn M, Letunic I, Jensen L J, et al. The SIDER database of drugs and side effects[J]. Nucleic Acids Research, 2016, 44(d1):D1075-D1079.

[15] Christoph Steinbeck , †, Yongquan Han †, Stefan Kuhn †, et al. The Chemistry Development Kit (CDK): An Open-Source Java Library for Chemo- and Bioinformatics[J]. ChemInform, 2003, 43(21):493-500.
[16] Burbidge R, Trotter M, Buxton B, et al. Drug design by machine learning: support vector machines for pharmaceutical data analysis.[J]. Computers & Chemistry, 2001, 26(1):5-14.
[17] Byvatov E, Schneider G. Support vector machine applications in bioinformatics.[J]. Applied Bioinformatics, 2003, 2(2):67-77.
[18] Peterson L. K-nearest neighbor.[J]. Scholarpedia, 2009, 4(2).