ISSN: 2454-9940

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

E-Mail :
editor.ijasem@gmail.com
editor@ijasem.org

www.ijasem.org

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN2454-9940

www.ijasem.org

Vol 18, Issue 2, 2024

Deep Side: A Deep Learning Framework for Drug Side Effect Prediction

[1] A.N. RAMAMANI, [2] CHINDADA VIJAYA

[1](Assistant Professor), MCA, S.V.K.P & Dr K.S. Raju Arts & Science College

[2]MCA, scholar, S.V.K.P & Dr K.S. Raju Arts & Science College

ABSTRACT

Drug failures due to unforeseen adverse effects at clinical trials pose health risks for the participants and lead to substantial financial losses. Side effect prediction algorithms have the potential to guide the drug design process. LINCS L1000 dataset provides a vast resource of cell line gene expression data perturbed by different drugs and creates a knowledge base for context specific features. The state-of-the-art approach that aims at using context specific information relies on only the highquality experiments in LINCS L1000 and discards a large portion of the experiments. In this study, our goal is to boost the prediction performance by utilizing this data to its full extent. We experiment with 5 deep learning architectures. We find that a multi-modal architecture produces the best predictive performance among multi-layer perceptron-based architectures when drug chemical structure (CS), and the full set of drug perturbed gene expression profiles (GEX) are used as modalities. Overall, we observe that the CS is more informative than the GEX. A convolutional neural network-based model that uses only SMILES string representation of the drugs achieves the best results and provides 13.0% macro-AUC and 3.1% micro-AUC improvements over the state-of-the-art. We also show that the model is able to predict side effect-drug pairs that are reported in the literature but was missing in the ground truth side effect dataset.

1.INTRODUCTION

Computational methods hold great promise for mitigating the health and financial risks of drug development by predicting possible

side effects before entering into the clinical trials. Several learning based methods have been proposed for predicting the side effects of drugs based on various features such as: chemical structures of drugs [25, 1, 23, 8, 19, 34, 17, 9, 2, 5], drug-protein interactions [35, 33, 8, 19, 34, 17, 37, 2, 15, 36], protein-protein interactions (PPI) [8, 9], activity in metabolic networks [38, 26], pathways, phenotype information and gene annotations [8]. In parallel to the above mentioned approaches, recently, deep learning models have been employed to predict side effects: (i) [31] uses biological, chemical and semantic information on drugs in addition to clinical notes and case reports and (ii) [4] uses various chemical fingerprints extracted using deep architectures to compare the side effect prediction performance.

While these methods have proven useful for predicting adverse drug reactions (ADRs – used\ interchangeably with drug side effects), the features they use are solely based on external knowledge about the drugs (i.e., drug-protein interactions, etc.) and are not cell or condition (i.e., dosage) specific. To address this issue ,Wang et al. (2016) utilize the data from the LINCS L1000 project [32]. This project profiles gene expression changes in numerous human cell lines after treating them with a large number of drugs and small-molecule compounds. By using the gene expression profiles of the treated cells, [32] provides the first comprehensive, unbiased, and cost-effective prediction of ADRs. The paper formulates the problem as a multi-label classification task. Their results suggest that the gene expression profiles provide context-dependent information for the side-effect prediction task. While the LINCS dataset contains a total of 473,647 experiments for 20,338 compounds, their method utilizes only the highest quality experiment for each drug to minimize noise. This means that most of the expression data are left unused, suggesting a potential room for improvement in the prediction performance. Moreover, their framework performs feature engineering by transforming gene expression features to enrichment vectors of biological terms. In this work, we investigate whether the incorporation of gene expression data along with the drug structure data can be leveraged better in a deep learning framework without the need for feature engineering.

In this study, we propose a deep learning framework, Deep Side, for ADR prediction. Deep Side uses only (i) in vitro gene expression profiling experiments (GEX) and their experimental meta data (i.e., cell line and dosage - META), and (ii) the chemical structure of the compounds (CS). Our models train on the full LINCS L1000 dataset and use the SIDER dataset as the ground truth for drug - ADR pair labels [13]. We experiment with five architectures: (i) a multi-layer perceptron (MLP), (ii) MLP with residual connections (Res MLP), (iii) multi-modal neural net- works (MMNN. Concat and MMNN. Sum), (iv) multi-task neural network (MTNN), and finally, (v) SMILES convolutional neural network (SMILES Conv). We present an extensive evaluation of the above-mentioned architectures and investigate the contribution of different features. Our experiments show that CS is a robust predictor of side effects. The base MLP model, which uses CS features as input, produces _11% macro-AUC and _2% micro- AUC improvement over the state-of-the-art results provided in [32], which uses both GEX (high quality) and CS features. The multi-modal neural network model, which uses CS, GEX and META features and uses summation in the fusion layer (MMNN. Sum) achieves 0:79 macro-AUC and 0:877 micro-AUC which is the best result among MLP based approaches. We also find out that when the chemical structure features are fully utilized in a complex model like ours, it overpowers the information that is obtained from the GEX dataset. The convolutional neural network that only uses the SMILES string representation of the drug structures achieves the best result among all the proposed architectures with provides 13:0% macro-AUC and 3:1% micro-AUC improvement over the state-of-the-art algorithm. Finally, inspecting the confident false positives predictions reveal side effects that are not reported in the ground truth dataset, but are indeed reported in the literature.

2.LITERATURE SURVEY

The authors report that their best result is obtained with the feature set that is a combination of gene ontology (GO) transformed gene expression profiles and chemical structures (CS). Their set of drugs with this feature set (GO + CS) contains 791 compounds. We use these 791 drugs to build

our models. In total, there are 18,832 experiments for these 791 drugs in the LINCS L1000 dataset. The META information for each of the 18,832 experiments from the LINCS project is also used as features. META information contains (i) the cell line on which the experiment is conducted on, (ii) the timing of the experiment, and (iii) dosage information. The meta information exists for 70 cell lines, 20 dosage levels and 3 time points (i.e. 6h, 24h, 48h). Note that for a given drug, the experiments do not cover all possible combinations of these conditions. META data is represented as one-hot encoding vectors. The corresponding feature vector has a length of 93. The total length of the concatenated GEX and META feature vectors is 1071. For all models, whenever META data is used, it is concatenated with the 978 landmark GEX features. We obtain the drug side effect information (labels) from the SIDER Database [13] (downloaded on Feb 5, 2018). The side effects that are observed with fewer than ten drugs are excluded as also done in [32]. This filtering stage leaves us with 1052 side effects in total. In order to group side effects, we utilize the ADR ontology database (ADReCS), which provides a hierarchical classification of side effects in a four-level tree [3]. The CS features are encoded with OpenBabel Chemistry Toolbox [20] to create a 166-bit MACCS chemical fingerprint matrix for each drug (a binary vector of length 166). A SMILES string is an alternative representation for the 2D molecular graph of a drug/small molecule as a 1D string. The SMILES strings are downloaded from PubChem [11]. These are used to create the chemical fingerprints of the drugs for the 1D convolution used in SMILESConv model. RDKit Cheminformatics toolbox is used to extract extended SMILES Strings of the drugs [14]. The extended SMILES strings contain all the primary chemical bonds as well as the hydrogen bonding information explicitly. Zero-padding is used to have a uniform representation among all drugs. The alphabet contains 33 unique characters, including the end of sequence character. We further generate a pruned drug dataset to compare SMILESConv model with others. We filter out drugs with SMILES representation that have less than 100 characters and more than 400 characters. 615 out of 791 drugs pass this filtering step.

For these drugs, we apply the additional filtering for removing side effects with less than ten drugs. In the end, 615 drugs and 1042 side effects pairs remain in this pruned dataset. Finally, we remove the characters that occur only once in all SMILES strings from the character vocabulary and replace them with underscore symbol.

3.SYSTEM ANALYSIS

A drug-drug interaction (DDI) is defined as an association between two drugs where the pharmacological effects of a drug are influenced by another drug. Positive DDIs can usually improve the therapeutic effects of patients, but negative DDIs cause the major cause of adverse drug reactions and even result in the drug withdrawal from the market and the patient death. Therefore, identifying DDIs has become a key component of the drug development and disease treatment.

In this study, an existing system, develops a method to predict DDIs based on the integrated similarity and semi-supervised learning (DDI-IS-SL). DDI-IS-SL integrates the drug chemical, biological and phenotype data to calculate the feature similarity of drugs with the cosine similarity method. The Gaussian Interaction Profile kernel similarity of drugs is also calculated based on known DDIs. A semi-supervised learning method (the Regularized Least Squares classifier) is used to calculate the interaction possibility scores of drug-drug pairs. In terms of the 5-fold cross validation, 10-fold cross validation and denovo drug validation, DDI-IS-SL can achieve the better prediction performance than other comparative methods. In addition, the average computation time of DDI-IS-SL is shorter than that of other comparative methods. Finally, case studies further demonstrate the performance of DDI-IS-SL in practical applications.

Disadvantages

• The complexity of data: Most of the existing machine learning models must be able to accurately interpret large and complex datasets to detect an accurate Drug Side Effect.

• Data availability: Most machine learning models require large amounts of data to create accurate predictions. If data is unavailable in sufficient quantities, then model accuracy may suffer.

• Incorrect labeling: The existing machine learning models are only as accurate as the

data trained using the input dataset. If the data has been incorrectly labeled, the model cannot make accurate predictions.

Proposed System

Multi-layer perceptron (MLP) Our MLP [22] model takes the concatenation of all input vectors and applies a series of fully-connected (FC) layers. Each FC layer is followed by a batch normalization layer [10]. We use ReLU activation [16], and dropout regularization [27] with a drop probability of 0:2. The sigmoid activation function is applied to the final layer outputs, which yields the ADR prediction probabilities. The loss function is defined as the sum of negative log-probabilities over ADR classes, i.e. the multi-label binary cross-entropy loss (BCE). An illustration of the architecture for CS and GEX features is given in this system.

Residual multi-layer perceptron (ResMLP) The residual multi-layer perceptron (ResMLP) architecture is very similar to MLP, except that it uses residual-connections across the fully- connected layers. More specifically, the input of each intermediate layer is element-wise added to its output, before getting processed by the next layer. Such residual connections have been shown to reduce the vanishing gradient problem to a large extend [7].

This effectively allows deeper architectures, therefore, potentially learning more complex and parameter-efficient feature extractors. Multi-modal neural networks (MMNN) The multi-modal neural network approach contains distinct MLP sub-networks where each one extract features from one data modality only. The outputs of these sub-networks are then fused and fed to the classification block. For feature fusion, we consider two strategies: concatenation and summation. While the former one concatenates the domain-specific feature vectors to a larger one, the latter one performs element-wise summation. By definition, for summation based fusion, the domain-specific feature extraction sub-networks have to be designed to produce vectors of equivalent sizes. We refer to the concatenation and summation based MMNN networks as MMNN.Concat and MMNN.Sum, respectively.

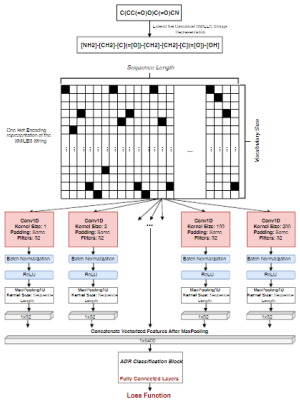
Multi-task neural network (MTNN) our multitask learning (MTL) based architecture aims to take the side effect groups obtained from the taxonomy of ADReCS into

account. For this purpose, the approach defines shared and task-specific MLP sub-network blocks. The shared block takes the concatenation of GEX and CS features as input and outputs a joint embedding. Each task-specific sub-network then converts the joint embedding into a vector of binary prediction scores for a set of inter-related side-effect classes.

Advantages

- The proposed system implemented many ml classifies for testing and training on datasets.
- The proposed system developed Convolutional neural networks (CNN) which are known to provide a powerful way of automatically learning complex features in vision tasks to find an accurate accuracy on the datasets.

4. OUTPUTSCREENS



5. CONCLUSION

The pharmaceutical drug development process is a long and demanding process. Unforeseen ADRs that arise at the drug development process can suspend or restart the whole development pipeline. Therefore, the a priori prediction of the side effects of the drug at the design phase is critical. In our Deep Side framework, we use context-related (gene expression) features along with the chemical structure to predict ADRs to account for conditions such as dosing, time interval, and cell line. The proposed MMNN model uses GEX and CS as combined features and achieves better accuracy performance compared to the models that only use the chemical structure (CS) finger-prints. The reported accuracy is noteworthy considering that we are only trying to estimate the condition-independent side effects. Finally, SMILES Conv model outperforms all other approaches by applying convolution on SMILES representation of drug chemical structure

6. REFERENCE

1. Atias, N., Sharan, R.: An algorithmic framework for predicting side effects of

drugs. Journal of Computational Biology 18(3), 207–218 (2011)

2. Bresso, E., Grisoni, R., Marchetti, G., Karaboga, A.S., Souchet, M., Devignes, M.D., Sma¨ıl-Tabbone, M.: Integrative relational machine-learning for understanding drug side-effect profiles. BMC bioinformatics 14(1), 207 (2013)

3. Cai, M.C., Xu, Q., Pan, Y.J., Pan, W., Ji, N., Li, Y.B., Jin, H.J., Liu, K., Ji, Z.L.: Adrecs: an ontology database for aiding standardization and hierarchical classification of adverse drug reaction terms. Nucleic acids research 43(D1), D907–D913 (2014)

4. Dey, S., Luo, H., Fokoue, A., Hu, J., Zhang, P.: Predicting adverse drug reactions through interpretable deep learning framework. BMC Bioinformatics 19 (12 2018). https://doi.org/10.1186/s12859-018-2544-0

5. Dimitri, G.M., Li´o, P.: Drugclust: A machine learning approach for drugs side effects prediction. Computational Biology and Chemistry 68, 204 – 210 (2017). https://doi.org/https://doi.org/10.1016/j.compbiolchem.2017.03.008