GT CSE 6242 Data and Visual Analytics Project Poster (2023 Spring)

Drug-Interaction Studies for a Side-Effect Alert System

Team064: Syed Shahbaz Gardezi, Yuzi Zhang, Charlie H. Gu, Kalpana Baheti

INTRODUCTION

Prescription of multiple medications at a time is fairly common. However, prediction of possible resulting side effects is a complex problem that involves structural similarity of drug pairs, drug environments, catalytic reactions, and so on. In this prototype, we explore the exact granularity of featurization required to reach a significant inference accuracy on our strictly drug-pair and side-effect data, with a suitable model. Please keep in mind that with this strict limit of data, full information on the relationship between drug-pair and side effect cannot be determined, hence the accuracy too will be capped. This project excavates useful inference for better drug-recommenders in future.

PROBLEM DEFINITION

The goal of this project is to be able to predict the most probable side-effect and trustworthiness for each pair of drugs entered based on selected featurisation of the drug molecules and classification algorithm based on these features. An interaction score will accompany the prediction. This problem is solved through the use of a webpage interface designed for easy entry of drugs and display of drug-pair side effects and trustworthiness of prediction.

METHODS

System Diagram:

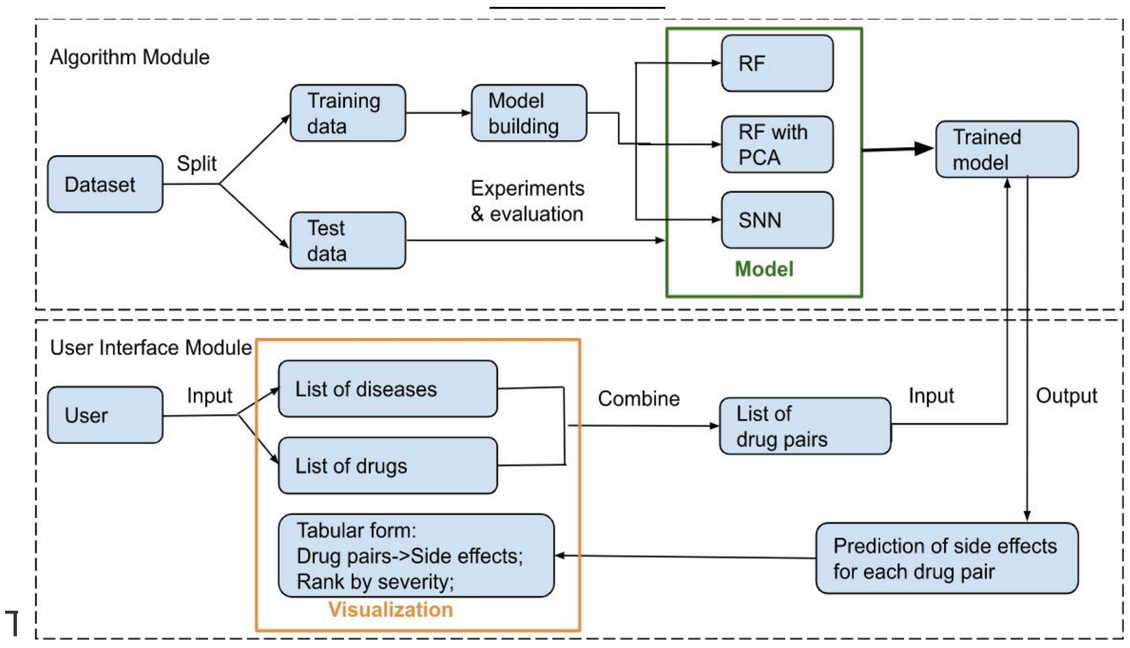

Figure 1. Diagram of the system.

There are two parts involved in the inference of this problem - the feature extraction (which is the hardest part), the algorithm used for fitting a relationship over these features and the target side-effect.

1. Decomposing molecule into common structures.
2. **Representing molecule in terms of residual atomic charges.**
3. Representing in spatially-contextual residual atomic charges.

Featurization method 3, is a highly sophisticated method (using Coulomb matrices that are adjusted using eigenvectors to incorporate spatial relevance) - but takes exceeding computation power, hence method 2 was adopted. The two algorithms tested were **random forest,** and **sequential neural network**. In each, a variant with **Principal Component Analysis** was tested. The best performing model was a post-PCA sequential neural network over 500 epochs of training, with an accuracy of 50%. Interaction scores were generated using a pre-trained ChemBERTa transformer with a sigmoid decision-maker.

DATA

The dataset '**TWOSIDES' from TDC** contains drug-drug interactions.We focused on 10 side effects out of 1000, which amounts to over 54000 drug pairs on training dataset and 23000 drug pairs on test dataset.

Data preprocessing was performed prior to model building. **Processed data in total amounted to 380 MB. This contained 730 features, size of each drug vector being 315 values of residual atomic charges.**

Dataset	Description
TDC TWOSIDES	Drug Pairs and Side Effects

Table 1. Dataset is available to download from Therapeutics Data Commons.

EXPERIMENTS AND RESULTS

Approaches were assessed on raw accuracy score determined by performing a cross validation test split.

1. **A sequential neural network for the residual charge relationship** seemed a good fit since the weighted sum of products continually updated towards minimum loss fits the nature of how the atomic charges contribute in synergy.
2. **Preprocessing with PCA** assisted with extracting most valuable information from the features.
3. **The sequential neural network ran over 500 epochs**, and had the following specifications -
   A. Input dimensions = 730
   B. 128 x 64 with 'ReLU' activation to enforce positive values.
   C. 10 neurons for softmax to get probabilistic accuracy.
   D. Sparse Categorical Cross-Entropy for loss calculation with Adam Optimizer
4. Accuracy tapered at around 500 epochs.

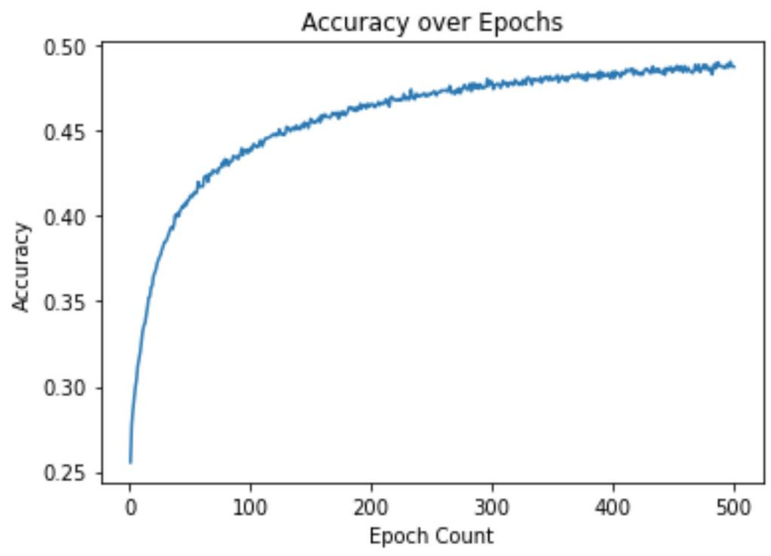

Figure 2. Stabilizing of accuracy over training

RESULTS:

1. 500 EPOCHS - Training: 53%, Testing: 15%
2. 100 EPOCHS - Training: 40%, Testing: 20%
3. 50 EPOCHS - Training: 30%, testing 30%

Highest accuracy: 53% on Training.

Visualisation of output with interaction score by ChemBERTa.

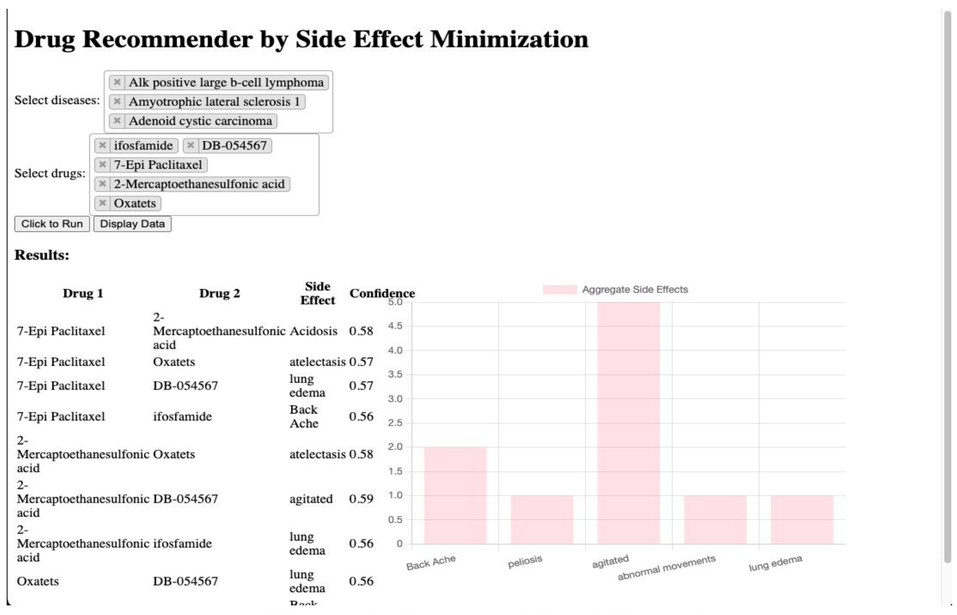

Figure 3. Webpage Visual Example.

In comparison with other methods, there does not seem to be a lot of literature on inferring from drug-pairs alone and relating to side-effects. Even drug interaction checkers have a lot more data regarding environment. Given our scope and deductive approach of starting from higher granularity of features and working our way to more fundamental aspects of molecules that causes interactions issues, and narrowing in on a logical algorithm to use within computation power, this would be an efficient PoC and a great starting point for incorporating further information.