MTA |The Academic College of Tel Aviv-Yaffo | ML Methods | George Kour
Student: Avlas Kfir
ID: 060519071

Drug repurposing for the SARS-CoV-2 pandemic - predicting whether an existing drug will be tested in clinical trials.

20th April 2023

Project Proposal

Abstract

The active global SARS-CoV-2 pandemic caused more than 763 million cases and 6.9 million deaths worldwide. The development of completely new drugs for such a novel disease is a challenging, time intensive process. This emphasizes the importance of drug repurposing, where treatments are found among existing drugs that are meant for different diseases.

A common approach to this is based on knowledge graphs that condense relationships between entities like drugs, diseases and genes. Graph neural networks (GNNs) can then be used by predicting links in such knowledge graphs. Expanding on state-of-the-art GNN research, Doshi et al. developed thes Dr-Covid model . Extension of their work presented by Sarel Cohen et al. by using additional output interpretation strategies. The research uses the concept of graph embeddings, which map fixed-size feature vectors to graph nodes and relations based on deep neural networks (DNNs).

My project will be based on Sarel Cohen et al. research. The nodes and relations graph represented by a matrix is the input for my project. By adding information from clinical trials as output target data, I will use a machine learning algorithm to train a classifier model and predict whether an unseen drug will be tested in COVID-19-related clinical trials.

Introduction

BACKGROUND

With the novel coronavirus, a global pandemic with serious socio-economic implications for most parts of our daily lives is active [1]. The limited ability to take precautions for an unsuspected event like this and the rapid spread make finding an effective treatment as necessary as difficult, since the disease-specific knowledge is limited at the beginning and human lives are lost every day. Known and approved drugs happen to be well-studied, thus, they pose a good starting point for swift development of treatments, and an emerging tactic in fighting the pandemic. DrugBank, an extensive database compiling information about drugs approved by the US Food and Drug Administration as well as experimental drugs, contained more than 2300 approved drugs and over 4500 experimental drugs as of 2018; both with a strong upward trend. This emphasizes the need for computer aided development of treatments.

CURRENT APPROACHES

Drug repurposing with knowledge graphs, is the current state-of-the-art approach for finding possible treatments for novel diseases among known drugs using machine learning. Applying drug repurposing allows for a better way to maneuver through the pandemic. It can lead to better treatments for patients infected with one of the COVID-19 strains and a better understanding of the characteristics of the individual strains. Expanding on state-of-the-art GNN research, Doshi et al. developed thes DR-Covid model [2] . Extension of their work presented by Sarel Cohen et al. [3] by using additional output interpretation strategies. The research approaches the problem of drug repurposing using machine learning, focusing on deep learning methods for predicting unknown links between entities in a knowledge graph.

PROBLEM DEFINITION

The research presents different strategies for interpreting the scores that the model outputs for the application of predicting the top-*r* most promising drug nodes for a given set of COVID-19 disease nodes, and yields a matrix of scores, then an aggregation strategy takes the matrix of standardized scores and derives a list of drugs from it. The top-*r* of which are the results. Finally there is a check of how many of the drugs were or are in clinical trials. There are several aggregation strategies which can lead to a different final drug list and the challenge is to decide which of the strategies may lead to better results.

Today, after the disease has already been studied, we have a partial knowledge about effective drug treatment and about drugs tested in clinical trials. For this reason, my project suggests another way of using the previous research scores. After getting the matrix, instead of deriving a list of drugs from it (by some aggregation strategy) I will use the score as an input feature, and check which of the drugs have been actually found effective or tested in clinical trials. This way I will be able to use a supervised machine learning model to predict whether an existing drug will be tested in COVID-19-related clinical trials.

EXPECTED CHALLENGES

Imbalanced data

Typically refers to a problem with classification models where the classes are not represented equally - classification data set with skewed class proportions. Imbalanced dataset may cause frustrating results when the classification model may get more than 90% accuracy immediately which turned out to be a lie. One of the challenges in the project will be handling imbalanced data. The dataset is expected to be imbalanced in a way that a major part of it will belong to one class. The vast majority of existing drugs are meant for different diseases and not expected to be tested in a COVID-19-related clinical trials. The vast majority of the transactions expected to be in the “Not-Tested” class and a very small minority expected to be in the “Tested” class. With so few positives relative to negatives, the training model will spend most of its time on negative examples and not learn enough from positive ones.

Feature selection technique

The desirable dataset for the project will contain all the details that exist for disease and drug. For this reason the dataset is expected to contain multiple features which is relatively high. it’s almost rare that all the variables in the dataset are useful for building a model. Adding redundant variables reduces the model’s generalization capability, may cause overfitting, and may also reduce the overall accuracy of a classifier. Furthermore, adding more variables to a model increases the overall complexity of the model.

Work Plan

METHOD

Supervised learning

Algorithms build a mathematical model of a set of data that contains both the inputs and the desired outputs. The data consists of a set of training examples. Each training example has one or more inputs and the desired output. Through iterative optimization of an objective function, supervised learning algorithms learn a function that can be used to predict the output associated with new inputs [4]. I will use a supervised machine learning classification algorithm to train a model on a given dataset. The input of the model dataset is a knowledge graph given by Sarel Cohen *et al*. research [3]. The goal of the project is to predict which of the drugs, given their prediction score to treat a COVID-19 variant, are likely to be tested in a COVID-19-related clinical trials.

Learning supervised problem

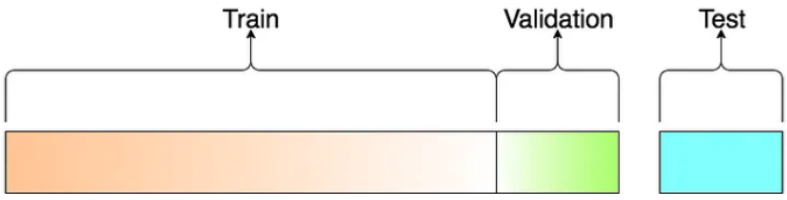
Given a set of n training examples $S = \{ (x(i), y(i)) \}$ $1 <= i <= n$ such that $x(i)$ = feature vector of i example, $y(i)$ is the target value (label) of i example. Assuming there exists an unknown oracle function $f: X - Y$ that return the correct output for any input, a learning algorithm seeks a function $h: X -> Y$ such that $f(x) \sim h(x)$.

DATESET

Sarel Cohen et al. work [3] relies on the Drug Repurposing Knowledge Graph (DRKG), which compiles data from different biomedical databases and uses 98 edge types between 4 entity types, namely gene, compound, anatomy and disease. In particular, it contains drugs and substances as *compound* entities, as well as different COVID-19 variants as *disease* entities. Finally there are 5000 drug entities and 33 different COVID-19 entities. The edge types include *compound-treats-disease* edges, which is the kind of edge the model predicts.This is the input for my project.

DATASET SPLIT

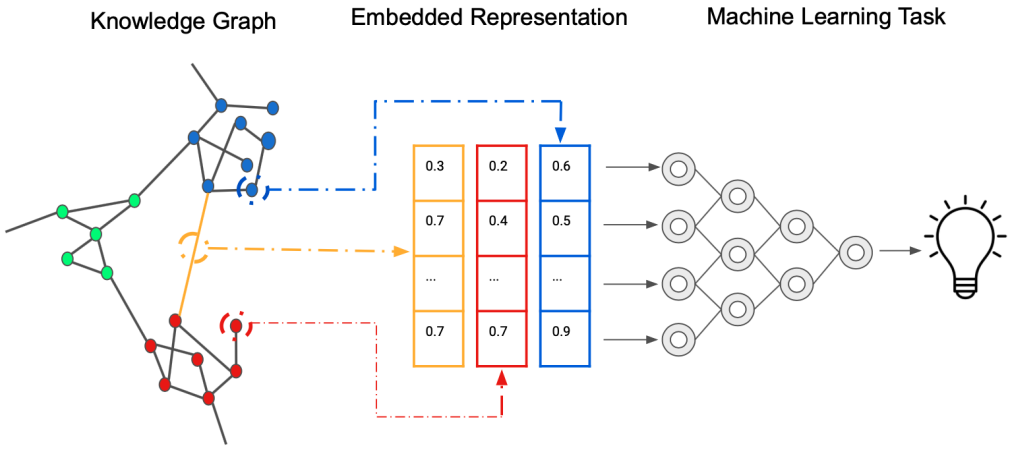
In general, *Training* dataset *is* the sample of data used to fit the model, *Validation* dataset is the sample of data used to provide an unbiased evaluation of a model fit on the training dataset while tuning model hyperparameters and *Test* dataset is the sample of data used to provide an unbiased evaluation of a final model fit on the training dataset.



The *Test* dataset provides the gold standard used to evaluate the model. It is only used once a model is completely trained (using the train and validation sets).

DATA PREPROCESSING

The data is based on the Drug Repurposing Knowledge Graph (DRKG), which compiles data from different biomedical databases. Next step is creating neighborhood *graph embeddings [5]*, which map fixed-size feature vectors to graph nodes and relations. Finally we get 5000 drug entities and 33 different COVID-19 entities. The edge types include *compound-treats-disease* edges, which is the kind of edge the model predicts. The nodes and relations graph represented by a matrix is the input to my machine learning task.



ALGORITHM

The input graph is represented by a two-dimensional matrix. *d*ifferent COVID-19 variants as *disease* vs drugs approved by the US Food and Drug Administration as *drugs*. Every cell i,j in the matrix has a value which is the prediction score of a disease j to be treated by a drug i. The *disease* feature , as well as the *drug* feature*,* is actually a vector consisting of multiple features. All of them will be a part of the learning model. In addition, as the disease has already been studied, the dataset will contain *is_drug_effective* feature and an output target variable showing whether a disease was tested in COVID-19-related clinical trials. Finally my dataset will consist of multiple labeled input features (of *disease , drug, score*) and a target binary variable (*is_tested_in_covid_clinical_trials*) . I will use a supervised machine learning classifier to train and predict whether an existing drug will be tested in COVID-19-related clinical trials. I will use the technique of dataset splitting in order to train and validate the results. Respectively to the challenges mentioned above, I will have to choose the most suitable solution and appropriate algorithm to deal with multiple features and imbalanced data as well as a loss function for evaluating the model and its accuracy.

References

1. World Health Organization | COVID-19 Dashboard

2. Doshi S, Chepuri SP. Dr-COVID: Graph Neural Networks for SARS-CoV-2 Drug Repurposing. CoRR. 2020.

3. Cohen S, Hershcovitch M, Taraz M, Kißig O, Issac D, Wood A, et al. (2023) Improved and optimized drug repurposing for the SARS-CoV-2 pandemic. PLoS ONE 18(3): e0266572. https://doi.org/10.1371/journal.pone.0266572

4. Machine learning | Wikipedia

5. Knowledge graph embedding