Machine Learning Project Proposal: Mechanisms of Action (MoA) Prediction

Malte Meng, 2020280441 (Leader)

Khang Hui Chua, 2020280442

Kai Wen Yoke, 2020280598

Abstract

This proposal introduces the Kaggle competition we have chosen: MoA prediction of drug sample given gene expression and cell viability data. We briefly give an overview of the biological context of the task, describe the dataset and potential difficulties, and suggest some promising ideas from related work by other Kaggle competitors that we can try out later on.

1 Topic

For our Machine Learning project, we have decided to choose a currently active Kaggle competition "Mechanisms of Action (MoA) Prediction" that ends on 30th November: `https://www.kaggle.com/c/lish-moa/`.

1.1 Problem Statement

Given data on gene expression and cell viability levels of cell samples in response to being treated by a drug sample at a particular dosage and for a particular time duration, the task is to predict the MoA of the drug. This is a multi-label classification task, as each drug sample can be associated with more than one MoA. The accuracy of solutions will be evaluated on the average value of the logarithmic loss function applied to each drug-MoA pair.

1.2 Dataset overview

We have 23814 data samples in the training set. The entire dataset includes roughly 5000 unique drugs, with each drug profiled multiple times at different dosages and treatment times. There are 772 **gene expression features**, with each gene feature representing the expression level of a particular gene, and 100 **cell viability features**, with each cell viability feature representing the viability of one cell line, meaning that there are 100 cell lines in this dataset. The actual names of the gene features and cell types are hidden, so a hand-crafted approach would not be possible. The **cp_type** attribute indicates whether a data sample is a drug or a control, and a control sample would have no associated MoAs, but note that there are data samples which are not controls and also do not belong to any of the 206 MoAs that are our target labels tested during scoring. However, the dataset also contains an additional non-scored 402 target MoAs for each data sample. The **cp_dose** is a binary feature and comprises two levels of dosage D1 or D2, and **cp_time** is a categorical feature comprising three treatment durations (24, 48 or 72 hours). It should also be noted that the classes in the dataset are highly imbalanced as many of the target MoAs have only one entry each while some have hundreds.

Preprint. Under review.

2 Introduction

Understanding the biological mechanisms behind the workings of different drugs is essential for drug discovery. Discovering new drugs involve identifying protein targets associated with a disease, and developing a molecule that can act on that protein target to counter the disease. The biological activity of a given molecule is described and referred to as a mechanism-of-action (MoA). However, the MoA of a drug cannot be determined directly and one approach is to treat a sample of human cells with the drug and then measure the cellular responses such as gene expression and cell viability. When drug molecules act on cells such as by binding or inhibiting a membrane protein receptor, the cells will reduce or increase the expression of specific genes, and thus gene expression can be indicative of a drug action mechanism. Similarly, a drug such as an antibiotic could potentially be lethal to certain types of cells such as bacteria, and hence cell viability could also be indicative of drug mechanism. These cellular reactions are then compared with known patterns in large genomic databases, such as The Connectivity Map Project that this Kaggle competition is based on.

2.1 Context of dataset collection

In this competition, the data is collected by a new technology L1000 that measures simultaneously human cells' responses to drugs in a pool of 100 different cell types. This means that all 100 cell types are gathered in a sample, which are then administered with a drug with a particular dosage and treatment time. While cell viability is calculated by counting the number of healthy cells remaining for each cell line, gene expression data is calculated over all cell lines. This set of cellular response data is a small subset of what is usually collected and analysed to determine the MoA of the molecule, and the task is to use only this data to predict the MoAs associated with that drug molecule.

3 Related work and ideas

Much useful information has been discussed on the Kaggle competition discussion boards. Most of the leading strategies are based on neural networks.

3.1 Feature engineering

There is also much discussion on how features should be compactly represented (e.g. by PCA), by converting data attributes to 2D images and then applying CNNs (1), or by simply using the raw table format as input to a new neural architecture called TabNet (2). Another thing is roughly 40% of the data has no associated MoAs out of the scored 206 target MoAs (this includes control data samples and treatment data samples), but if we include the additional unscored 402 MoAs, the amount of data that has no associated MoAs decrease to roughly 23%. Therefore, it would be helpful if we can incorporate the additional unscored labels into training. It has been suggested that a new feature **non-score target** would help to distinguish between controls, drug samples with unscored MoAs and drug samples with scored MoAs. There is also the problem of how we can make use of the control data samples to improve classification accuracy, some have suggested using control data to normalise the drug data.

3.2 Training loss metric

Regarding training loss metric, there is also great variation as the competition loss metric (log-loss) is known to be a bad fit for an unbalanced problem like this, and instead Hamming loss, AUC, micro/macro-averaging and label smoothing (3) have been tried.

3.3 Data augmentation

Pseudo-labelling has always been popular in Kaggle competitions, and it might be useful to increase the size of the training set. Different amounts of noise can also be added to the training data to reduce overfitting and also augment the training set. Undersampling and oversampling certain classes could also be conducted to make the dataset more balanced. Perhaps, there can be some way to make use of external drug genome datasets to boost the training set.

3.4 Stratified K fold Cross Validation

Since the training dataset is small, it is important to do cross validation to average out the error for train-valid split for accurate evaluation.

3.5 Ensembling and stacking

This is always an important step in Kaggle competitions.

References

[1] B. Lyu and A. Haque, "Deep learning based tumor type classification using gene expression data," *bioRxiv*, 2018. [Online]. Available: https://www.biorxiv.org/content/early/2018/07/11/364323

[2] S. O. Arik and T. Pfister, "Tabnet: Attentive interpretable tabular learning," 2020.

[3] R. Müller, S. Kornblith, and G. Hinton, "When does label smoothing help?" 2020.