Machine Learning: Classifying Drug Mechanisms of Action Using Cell Viability and Gene Expression Data – Kaggle Competition

Sammy Lahrime
University of Virginia, class of 2020
Sl8rn@virginia.edu
Competition Link:
https://www.kaggle.com/c/lish-moa/overview

1. Preliminary Information

1.1 Motivation

In traditional drug research and development, scientists would opt for a top-down approach in discovery where drug candidates were first observed as known natural compounds. For instance, acetaminophen (brand name: Tylenol) was first discovered as a traditional remedy before it was commercially synthesized and produced for the general public in the late 1800s. Further, the drug was developed without any underlying intuition about its mechanism of action or pharmacology (Kaggle Overview). Although this was a convenient approach at the time, there were a lot of challenges associated with this model of drug discovery making it a very suboptimal approach compared to modern day techniques. For the most, traditional drug discovery relies on luck as there is no guarantee that an effective traditional remedy exists for all diseases, especially for diseases that require a very targeted/complex mechanism of action. However, due to recent technological advancements and a deeper understanding of the biological mechanisms of disease, scientists may now opt for the bottom-up approach, where a desired drug candidate may be developed based on the inspiration of the known biological process that governs the disease as well as the known mechanism of action of the drug candidate. That being said, a crucial component in that drug research and development process is having the ability to accurately determine a drug's mechanism of action in the first place. In the context of the Drug Mechanisms Kaggle Challenge, the problem statement presents 5000 drug samples with over 200 features and asks for a multi-label classification of every drug sample with the labels being the 200 possible MOAs.

1.2 Background

The challenge presents features and labels that utilize vocabulary terms and concepts in Biology. Therefore, to provide the necessary context for the data, this terminology must be understood before the implementation and data preprocessing steps. First, we may start by defining the associated features that are categorical in nature, those being cp_type, cp_time, and cp_dose. Cp_vehicle is a categorical variable that simply distinguishes where or not a drug sample is a placebo. In this case, the placebo was categorized as "ctrl_vehicle" while the active drugs were categorized as "cp_vehicle". Cp_time indicates the treatment duration of the drug in hours, incurring the possible categories of 24 hours, 48 hours, and 72 hours. Cp_dose indicates the dosage amounts categorically, incurring the only possible distinguishers of "high" and "low". As far as the numerical features, there are cell-viability features and gene expression features denoted as "c-" and "g-" respectively. Cell viability is a measure of the cell health after a drug treatment in terms of the proportion of living cells to the cell population post-treatment. There are 100 distinct cell viability features each of them indicating the cell viability of a distinct cell type. Gene expression is the biological process of interpreting genetic information to synthesize an associated genetic product (mostly proteins). The measure of gene expression is found through a well-known gene-expression profiling assay called the L1000. There are a total of 772 gene expression features, where each enumerated feature represents a distinct gene and the expression of that gene is measured as a numerical value according to the L1000. Absolute gene expression values (> 2 or < negative 2) represent the drug having a very significant effect on that particular gene, while values closer to 0 indicate that the drug had a very minimal effect on the gene, while an absolute value would indicate that a drug had a very significant effect. According to one of the competition hosts who goes by the Kaggle Username, "mrbhbs", he states, "We tried to strike a balance between providing sufficient domain knowledge and obfuscating information. The decision to obfuscate the names of the genes and cell lines was to encourage solutions that identify regularities specific of certain genes directly from the data (and therefore generalizable to other datasets) instead of incorporating prior knowledge." Therefore, this warrants that there is an intended degree of anonymity between features, and the sole purpose of the challenge is to identify key patterns rather than attempt to do further research as to the exact gene/cell-type a gene expression feature and cell-viability feature represents respectively. As mentioned previously, the drug samples are being classified by their mechanisms of action. Mechanism of Action (MoA) is formally defined as the resulting biochemical reactions of an administered drug by which it exerts its effects. However, distinguishing MoAs is categorical in nature, so short-hands are used as the label for corresponding MoAs. For instance, a drug being classified as an "aromatase inhibitor" would be it's corresponding MoAs. Further, drugs may have multiple MoAs which warrants the multi-label classification for the drug samples.

1.3 Related Work

The Mechanisms of Action (MoA) prediction challenge was hosted by the Laboratory for Innovation Science at Harvard, being launched on September 3rd, 2020, and closing on November 30th, 2020. Therefore, for the majority of the timeline of this project, it

was an active challenge with over 88,000 entries. A lot of competitors were transparent about the basic design choices and implementation of their entries, therefore there were a lot of available resources to reference for ideas and means of optimization. Primarily, there were two notebooks referenced that implemented Feed Forward Neural Networks to solve the challenge. The first notebook was submitted by the Kaggle user, "Amshra267". Their implementation trained a Neural Network, however they made changes in the optimizer, activation function, and neuron architecture. Specifically, the activation function was selected as the PRelu function. The primary difference between Relu and PRelu is that the PRelu accounts for the negative input cases, unlike ReLU, which will zero them out by default (Amanmishra4yearbtech). Furthermore, the notebook used a variation of the Adam optimizer that included weight decay. Weight decay is an approach to simplify this model to combat overfitting. Specifically it achieves this through regularization, where the features weighted the most highly are given priority. Consequently, the notebook was able to reduce the number of features from 785 to the top 696. Finally, dropout layers were added to further combat overfitting in the model, with dropout rates ranging from (0.2-0.41). The second notebook was submitted by the Kaggle user, "sarthak97". The approach consisted of a Neural Network model trained with 5 folds, extracting the top features, early-stopping, and the addition of dropout layers (Rana). Outside of the Kaggle competition, the amount of publications and studies that have attempted the same target task are limited and relatively recent. This may be attributed to the novelty of using cell-viability and gene expression data to classify MoAs instead of other features available. For instance, interestingly enough MoAs may be modelled through the use of Convolutional Neural Networks on image data features. A publication produced by the Cancer Research UK Edinburgh Centre reads, "The aim of the current study was to evaluate and compare the performance of a classic ensemble-based tree classifier trained on extracted morphological features and a deep learning classifier using convolutional neural networks (CNNs) trained directly on images from the same dataset to predict compound mechanism of action across a morphologically and genetically distinct cell panel" (Warchal).

1.4 Claim/Target Task

The task is a multi-label classification problem. Provided a total of 23,814 drug samples, each of which having 875 features, determine all of the corresponding MoA-labels (206 MoAs) of each drug sample. In the provided testing data, a sample should have a "1" for a MoA column if it has that MoA, and a "0" if it doesn't. Therefore, provided an input of 23,814 samples x 875 features, the output should be of the dimensions, 23,814 samples x 206 MoAs, where every entry denotes a "1" or "0". Further, accuracy may be determined by comparing the percentage of correct entries of the predicted output compared to the test output.

1.5 Proposed Solution

The proposed solution is to train a Feed Forward Neural Network using each sample and experiment with different optimization techniques and parameters. Each input will be fed into the input layer, with each sample consisting of its respective 875 features (1 x 875-dimension input). The output consists of a corresponding vector of classifications for all 206 MoAs, however all of the entries will be probabilities between 0 and 1, instead of the required "0" and "1" labels needed for a complete prediction. After these values are found, the probabilities that are greater than or equal to 0.5 will be mapped to a "1", while the other probabilities will be mapped to a "0". The model will initially be fitted with a .25 validation split, utilize a batch size of 100, and run through 100 epochs. Further, the Adam optimizer will be used along with a Binary Cross Entropy loss function. All of these parameters will be adjusted accordingly to increase the accuracy of the model.

2. Design & Implementation

2.1 Implementation

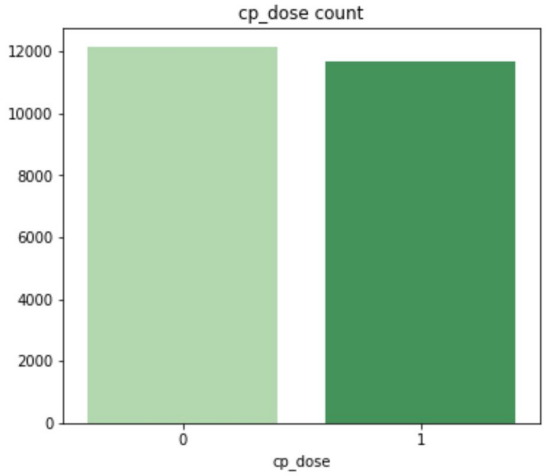
The first primary challenge was to correctly preprocess the data to ensure that it was in a purely numerical form for the training set. As outlined in section 1.2, there are three features that are categorical in nature. These features include cp_dose, cp_time, and cp_type. Consequently, these categories were mapped to values based on the amount of distinct possible values within each feature. First, cp_dose is known to incur only two possible values, "low" and "high". Consequently, "low" and "high" were mapped to 1 and 2 respectively. Further, cp_type was known to only incur two possible values, "ctrl_vehicle" (placebo) and "cp_vehicle", so they were mapped to 0 and 1 respectively. Finally, cp_time is known to incur only three possible values, "24 hours", "48 hours", and "72 hours". These possible values were mapped to 1, 2, and 3 respectively. In the process of training the Feed Forward Neural Network, there were a lot of considerations to make in parameter tuning and selection. The first primary challenge was identifying the nature of the Neural Network output. In this case, it is a multi-label classification problem. Consequently, a drug sample may belong to multiple different MoAs and it wouldn't be sufficient to classify each of them to a single mechanism. Therefore, the output layer must consist of all 206 of the possible MoAs. Afterwards, there had to be a decision made as to how many hidden layers the neural network would consist of. Initially, the model was fitted with only one hidden layer, then the accuracy was measured at the single layer as a baseline. Further, layers were increasingly added until the increased accuracy upon every added layer diminished to become insignificant. After this described iteration, it was found that a neural network of eight hidden layers produced the best accuracy without overfitting the model. Second, there were considerations in choosing the corresponding activation function. Upon the first trial, every layer utilized the softmax-activation function, however it wasn't expected that softmax would produce the best results for the problem at hand. After attempting the hidden layer activations functions of sigmoid, softmax, tanh, PReLU, and LeakyReLU, it was found that ReLU produced the best accuracy when implemented in the hidden layers. The next considerations were in selecting the hyperparameters for training the model, these being the batch size and epochs.

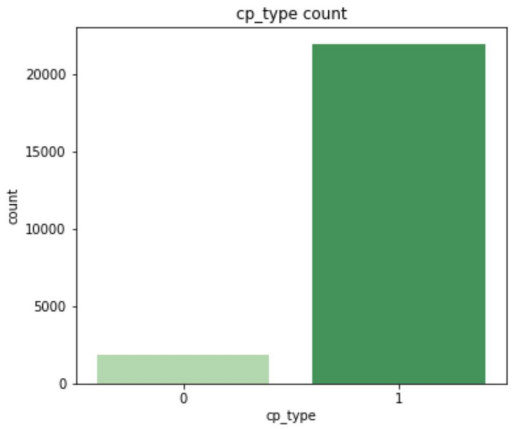
2.2 Data Summary

The data is composed of a provided testing and training set. The training set contains the 875 features (columns) for each drug sample, as well as a total of 23,814 drug samples (rows). Further the classification for each drug sample is also provided in the form of a table that is of the dimensions of 23,814 samples x 206 labelled MoAs. The drug sample is denoted as belonging to a certain MoA if it contains an entry value of "1" under the designated column of the MoA, and a "0" otherwise. The feature data is composed of 3 categorical features, 100 numerical cell-viability features (denoted as "c-"), and 772 numerical gene expression features(denoted as "g-"). The resulting prediction

should consist of an output of 23,814 samples x 206 labelled MoAs, with every entry filled with a “1” or “0”.

Figure 1: The frequency of each categorical feature, cp_dose, cp_type, and cp_time respectively..





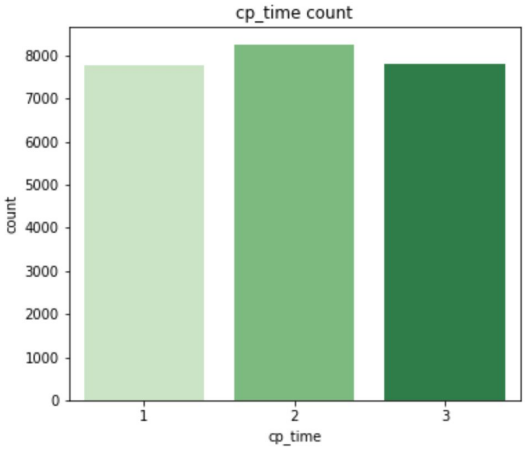


Figure 2: The relative frequencies of the cell-viability feature data.

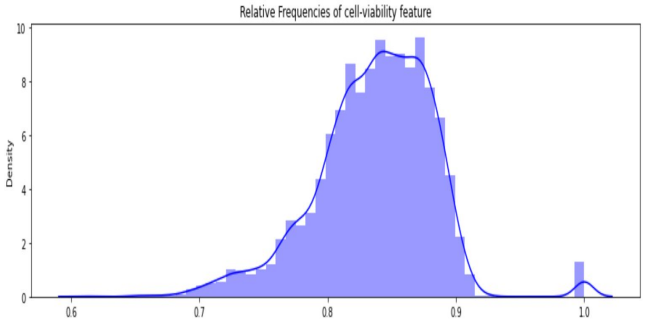
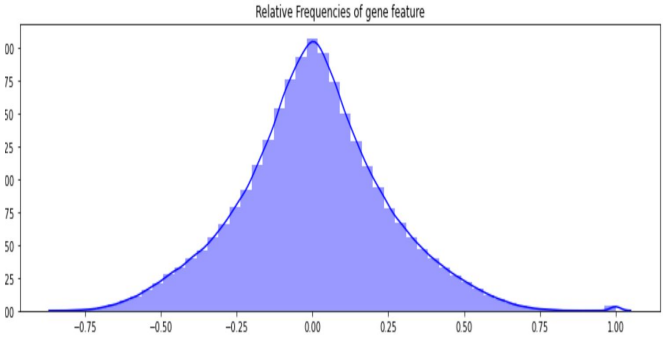


Figure 3 The relative frequencies of the gene expression feature data.



3. Results & Discussion

3.1 Experimental Results

After training the Neural Network, the most optimal hyperparameters were found to be 20 epochs, a batch size of 40, 8 hidden layers, a learning rate of 0.0001, and the use of ReLU loss function for each hidden layer respectively. Furthermore, it was found that the Square Hinge loss function yielded the highest accuracy out of all of the loss functions tried. Dropout layers with a dropout rate of 0.2 were also added in the Neural Network hidden layer. During the testing phase, it was found that the model yielded a final accuracy of about 43.64% with a 0.25 validation split.

3.2 Baselines

Formally, the competition assesses the notebook submissions with log-loss function instead of accuracy (the metric that this project used), as accuracy may lead to misleading results that are not indicative of prediction. That being said, the best public Feed Forward Neural notebook submissions utilize a very similar technique for mapping the categorical features, however it is common to not map the cp_time feature at all (24 hours, 48 hours, and 72 hours). Further, weight-decay optimizers, dropout layers, and some sort of k-fold validation are all pretty common amongst the public notebook submissions.

3.3 Conclusion

The objective of this project was to accurately predict the MoAs of each drug sample with the provided features using a Feed Forward Neural Network, then experimenting with different optimizations to improve the results. Although the resulting accuracy was by no means competitive relative to the more experienced submissions, it provided a lot of insight as to the methodology behind starting off with a provided dataset and task, then producing a catered pipeline to produce a model of the intended output. It was also a great learning experience with respect to the different optimization techniques that were attempted to improve the model. There is also a lot of excitement around the context of the task in general. As we develop better analytic techniques for observing Biological data, we will be able to concurrently build better models with greater predictive capabilities. Furthermore, as Machine Learning techniques advance, we will be able to draw more insight and produce better use cases.

4. REFERENCES

[1] Amanmishra4yearbtech, A. (2020, September 14). MoA : Keras Multilabel Neural Network v 2.0. https://www.kaggle.com/amanmishra4yearbtech/moa-keras-multilabel-neural-network-v-2-0

[2] Laboratory for Innovation Science at Harvard. (2020, September 3). Mechanisms of Action (MoA) Prediction. Retrieved December 13, 2020, from https://www.kaggle.com/c/lish-moa

[3] Rana, S. (2020, September 24). TF Keras : 5 Folds NN Starter🎯. https://www.kaggle.com/sarthak97/tf-keras-5-folds-nn-starter

[4] Scott J. Warchal, J. (2019, January 29). Evaluation of Machine Learning Classifiers to Predict Compound Mechanism of Action When Transferred across Distinct Cell Lines - Scott J. Warchal, John C. Dawson, Neil O. Carragher, 2019. https://journals.sagepub.com/doi/10.1177/2472555218820805

About the authors:

Sammy Lahrime is a third year student studying Computer Science in the School of Engineering & Applied Sciences of UVa.

Figure 4: Intuitive figure showing an interesting use case of MoA predictions for drug discovery: