CS 475/675 Project Proposal

Rupsa Acharya, Rishima Mukherjee, Harrison Khoo, Dimitri Lezcano
rachary6, rmukher9, hkhoo2, dlezcan1

Abstract

We propose a model to characterize small molecules as potential inhibitors of Epidermal Growth Factor Receptor (EGF), a protein upregulated in cancer cells. We will compare neural network and random forest models to determine which molecular fingerprints are good indicators of protein inhibition.

1 Project choice

Choose either a **methods** or **applications** project, and a subarea from the below table.

☒ **Applications**				
☒ Genomics data	☐ Healthcare data	☐ Text data	☐ Image data	☐ Finance data
☐ **Methods**				
☐ Fairness in ML	☐ Interpretable ML	☐ Graphical Models	☐ Robust ML	☐ Privacy in ML

2 Introduction

Epidermal Growth Factor Receptor (EGF) is a protein critical to cell division signalling pathways. EGF is often overexpressed in cancerous cells, resulting in rapid cell division and cancer proliferation. Thus, EGF is a common therapeutic target to slow down or stop cancer proliferation. The dataset we are working with distinguishes small molecules as inhibitors or not an inhibitor for eight signalling pathways with well known roles in cancer progression; we will focus on the approximately 6,000 samples that assess the molecules as inhibitors for the EGF kinase. The input for each molecule is 8192 binary features/fingerprints associated with each molecule. We will use two models to train and classify small molecules as inhibitors for the EGF pathway. Our primary model is a neural network and we shall use a secondary, random forest model as a comparison. We will output a prediction determining whether a specific molecule is an inhibitor to the EGF kinase.

3 Dataset and Features

The provided training and test datasets on Kaggle have 6055 and 882 small molecule inhibitor candidates, respectively, associated with EGFR [1]. For each small molecule, there is an inhibitor label and 8192 features/fingerprints that comprise the small molecule characteristics. As is, the small molecules and fingerprints are unlabeled. We will retrieve the molecule IDs using provided h5 files with similar data. The fingerprint IDs are nonsensical integers, so we will use a more standard numbering system in its place. Additionally, we hope to reduce the number of input features into the model, so we will perform principal component analysis to extract fingerprints of note.

	Label	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0

Figure 1: Example of small molecule fingerprints from Kaggle

4 Methods

We plan to use a feed-forward neural network to perform the classification of the molecular fingerprints to classify whether the molecule is a cancer inhibitor or not. Our model's hypothesis class will be the set of functional compositions of linear transforms followed by a non-linear activation (most likely ReLU). Our loss function will be Class-weighted focal cross-entropy as we have seen from the dataset that there exists a label imbalance between the non-inhibitors and inhibitors [2]. This loss allows for the rebalancing of weights associated to the loss of misclassified samples dependent upon the number of samples that are present in the training dataset. The optimization we plan on using is batch gradient descent using the Adam optimizer. Adam optimization is selected for its good performance on stochastic gradient descent, allowing for efficient optimization [3].

Furthermore, we plan to use random forest classification as an alternative learning method. The loss function we intend to use is Gini impurity as a measure of misclassification and will optimize our random forest classification by using a random sampled grid-search method over the parameters of the random forest classifier. The parameters that will be optimized will be: tree's max depth, tree's minimum samples to split, and the tree's minimum number of leaf nodes.

5 Deliverables

These are ordered by how important they are to the project and how thoroughly you have thought them through. You should be confident that your "must accomplish" deliverables are achievable; one or two should be completed by the time you turn in your Nov 19 progress report.

5.1 Must accomplish

1. Exploratory data analysis: feature and class distribution, correlations, etc.
2. Feature engineering and feature selection
3. Trained neural network model to predict effectiveness of small molecule better than random classification.
4. Compare neural network performance to random forest method.

5.2 Expect to accomplish

1. Achieve 75% accuracy with neural network model
2. Hyper-parameter tuning of neural network model for best performance
3. Compute neural network model performance metrics (accuracy, sensitivity, specificity, precision, recall, **F1 score**)

5.3 Would like to accomplish

1. Achieve 85% accuracy with neural network model
2. State of the art accuracy achieved.
3. Evaluate our neural network model on external, unlabeled PubChem dataset and validity of predictions with literature search

References

[1] "Cancer inhibitors kaggle dataset," https://www.kaggle.com/xiaotawkaggle/inhibitors, accessed: 2011-11-08.

[2] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, "Class-balanced loss based on effective number of samples," 2019.

[3] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2017.