Abu Qader

Computer Science Fair 2016

Within the subfield of cancers, there are thousands of treatments for the hundreds of cancers there are. Within these treatments, are millions of synthesized chemical compounds claiming to be the “fix” to cancer. Most medical professionals are faced with the daunting decision of personalizing a treatment plan tailored to that specific patient. They need it to be as effective, quick and efficient as possible. Cancer.org states that over 40% of cancer patients receive the wrong treatment that they need. Wrong treatment means wasted time that could’ve possibly saved the patient’s life, money that didn’t need to have been spent and a frantic search for a new treatment.

On the other side, the rise of genomics has proven to be successful. Genes are the coding blocks of the human system. Genes control the growth and decline of everything including physical and emotional traits. Mutated genes are also the cause of almost all tumors. Many cancer treatments aim towards blocking the mutated genes ability to function yet there are many compounding variables involved in the genome that may interfere with the effectiveness of the drug.

My proposed and implemented solution is to create a 2 stage system for big pharma companies and medical professional to test and predict the effectiveness of a drug. The first stage is meant for medical professionals. A super learner, also referred to as stacked boosting (See Part C on explanation), combined with genomic data of a patient can predict drug-gene interactions. This interaction helps predict the effectiveness of a certain type of chemogenetic therapy. The 2nd stage lies within big pharma companies that produce these drugs. By producing drugs quicker, we can help transform the field of smarter drugs. The second stage is a virtual screening tool that can (See Part C on explanation). By doing this, we can cut down on the animal testing stages of a drug and go straight to the drug experiments thus pushing more stable drugs to the public. Statistical machine-learning algorithms classify compounds as drug-like and nondrug-like based on their molecular descriptors.

A super learner or ensemble machine learning systems utilize multiple learning algorithms to obtain better predictive performance than could be obtained from any of the constituent learning algorithms. Some ensemble methods or highlighted below:

* Bootstrap aggregating, or bagging, is an ensemble method designed to improve the stability and accuracy of machine learning algorithms. It reduces variance and helps to avoid overfitting. Bagging is a special case of the model averaging approach and is relatively robust against noisy data and outliers.
* Boosting is an ensemble method designed to reduce bias and variance. A boosting algorithm iteratively learns weak classifiers and adds them to a final strong classifier.

After a weak learner is added, the data is reweighted: examples that are misclassified gain weight and examples that are classified correctly lose weight. Thus, future weak learners focus more on the examples that previous weak learners misclassified. This causes boosting methods to be not very robust to noisy data and outliers.

Both bagging and boosting are ensembles that take a collection of weak learners and forms a single, strong learner.

* Stacking is a broad class of algorithms that involves training a second-level "metalearner" to ensemble a group of base learners. The type of ensemble learning implemented is called "super learning", "stacked regression" or "stacking." Unlike bagging and boosting, the goal in stacking is to ensemble strong, diverse sets of learners together. By doing so, we can leverage the power of highly accurate models for even more accurate predictions.

The algorithms combined to create the super learner consisted of a deep learning neural network, a random forest algorithm, a gradient boosted algorithm, and a generalized linear machine.

* Deep learning neural network: Utilizes special nodes to create node-specific functions that an “input” runs through. Modeled after the human brain as it learns from example and utilizes around 100,000 epochs to fine tune.
* Random forest algorithm: Creates multiple decisions trees and trains them together. A random forest tree utilizes bagging by combining these decision trees and averaging their predictions.
* Gradient boosted algorithm: Creates really weak precision models and then combines them to form one very powerful model. It is another type of ensemble method. By boosting them, or normalizing the data, the models don’t have to worry about the saturation of the layers.
* Generalized linear machine: A form of linear regression. Linear regression plots the input data and creates a line of fit. A GLM does exactly that yet predictions and confidence ratings are based on how far away the prediction is from the nearest training point.

Chemo-genetics, a rising field in cancer medicine, focuses on the genetic aspect of cancer. It aims to fight the cause of cancer rather the symptoms. Drugs that fall under chemo-genetics aim towards attacking cell lines that contain mutated genes. These mutated genes cause the cells to become cancer cells. The inputs required for the 1st stage of the program (predictive drug-gene interaction) requires the following inputs:

* the drug that is considered to be used
* the cell line the medical staff focuses on attacking
* the number of mutations within that cell line
* the number cell lines with a coding sequence variant in the indicated gene
* Number of cell lines with a amplification in the indicated gene
* Number of cell lines with a homozygous deletion in the indicated gene

The program returns a 97% accuracy. The 1st stage utilizes a deep learning neural network paired with back propagation. It contains 2 hidden layers, the first consisting of 12 nodes and the second 8 nodes.

“Druglikeness” or the ability for a chemical to properly work within the body, is a set of rules a chemical must pass or it won’t actively work within the body. Virtual screening methods, which are fast, effective and comparatively cheap, can be used to evaluate these compounds in the early step of drug discovery and development studies. By utilizing virtual screening, we can cut down on costs and time thus helping release drugs quicker. Some qualities of a drug like substance are that they are both water and fat soluble and that they don’t contain conflicting chemical compounds. The inputs required for the second stage are

* Partition coefficient: the ratio between two “phases” such as the ratio between gas and liquid or solid and liquid
* polar surface area: the surface at the molecular level
* donor count: the amount of molecules willing to give an atom
* aliphatic ring count: number of those rings in the molecule that have non-aromatic bonds
* aromatic ring count: number of aromatic rings in the molecule
* balaban index: the number of edges in a molecular graph

The program returns a 98% accuracy. It utilizes a superlearner of the above algorithms.