Machine Learning Classification of Cancer Types

Ian Rolls and Josh Chen

Abstract

Cancer is the leading cause of death in the world; the application of machine learning models to classify cancer types based is a promising frontier that could aid in earlier detection and more accurate diagnosis. We applied several machine learning models to assess their accuracy in classifying cancer types using a dataset detailing patients' miRNA isoforms. We found that the SVM model was the most effective in the classification of cancer types. Potential future directions include developing models that could take in more robust datasets containing other biomarkers besides miRNA isoforms, as well as improved feature selection to continue improving the accuracy of our machine-learning models.

1 Introduction

Cancer is currently one of the most prominent health issues for the global human population. The random, mutagenic nature of cancer has made diagnosis especially difficult. Treatments are sometimes chosen on the basis of cancer origin, however, the ability of cancer to metastasize to other parts of the body from the primary site hinders the ability of medical professionals to make an accurate diagnosis without extensive tests. With the advancements of technology, researchers have identified numerous biomarkers that can assist with diagnosis. One such biomarker is miRNA. miRNA are non-coding, regulatory fragments of RNA. Different tissues express different isoforms of miRNA, and the same or similar types of cancer exhibit similar expression patterns (isoforms) (Lon). Thus, the implementation of machine learning models to characterize cancer types using miRNA isoforms can be a powerful tool in primary cancer identification.

Our study was inspired by Telonis' study on using presence/absence of miRNA isoforms to discriminate between 32 cancer types (Lopez-Rincon et al. 2020) Using the expression patterns of miRNA isoforms from patient data in TCGA (The Cancer Genome Atlas) database (Grossman et al. 2016), We seek to accurately classify 6 cancer types (breast invasive carcinoma, kidney renal clear cell carcinoma, lung adenocarcinoma, lung squamous cell carcinoma, pancreatic adenocarcinoma, uveal melanoma) using classification models such as KNNeighbors, stochastic gradient descent using a logistic loss function, support vector machine, and random forest classifiers. Then, we aim to compare our results to Telonis' study to identify any shortcomings of our models or improvements our models have made.

2 Dataset

The data set we use for training our models is taken from the journal article "Knowledge about the presence or absence of miRNA isoforms (isomiRs) can successfully discriminate amongst 32 TCGA cancer types" (Telonis et al. 2017). This dataset contains MicroRNA data for 2895 patients who are diagnosed with 6 different cancers. We assign each of these cancers a numerical label, which are listed in table 1.

An notable feature of the dataset is that the data for each cancer type contains a different number of observations. Table 1 shows the number of observations for each type of cancer.

Label	Cancer Type	Obs.
0	Breast Invasive Carcinoma	1096
1	Kidney Renal Clear Cell Carcinoma	544
2	Lung Adenocarcinoma	519
3	Lung Squamous Cell Carcinoma	478
4	Pancreatic Adenocarcinoma	178
5	Uveal Melanoma	80

Table 1: Label Catagories

Each patient file has data on the patient's MicroRNA. specifically, there is data on the presence and quantitiy of 1880 different miRNA strands in each patient. The metric used to determine the quantity of each miRNA type is reads per million miRNA mapped. Our target is the type of cancer and our features are the frequency of each type of miRNA.

3 Exploratory Analysis

We first performed a 3:1 train/test split on our data, giving us 1410 training examples and 470 test examples. As the distribution of labels in the dataset is not consistent, we stratify the data in our train/test split which preserves the distribution of labels. This is especially important for label 5 which has 80 total examples, as it would be easier for a majority of these to be randomized into the testing sample. Additionally we use the random_state feature in the sklearn package

to split the data identically which is necessary to compare results between different feature selection models.

We performed feature selection on the reads per million of the miRNA isoforms as the motivating paper suggests (Telonis et al. 2017). We binarized reads per million of miRNA isoforms, where 0 indicates absence and 1 indicates presence. Presence indicates that the reads per million fall in the top 20 percent of patients for the respective miRNA isoform. To identify whether our feature-selection positively impacts our models' performance, we compared the performance before and after feature-selection. We will call this model the binarized model, whereas our model without any feature selection will be called the basic model.

4 Experiments

We use 5 different classification techniques to find the best classification model for this problem. We are using K-Nearest-Neighbors (KNN), Support Vector Machines (SVM), Stochastic Gradient Descent (SGD), Random Forests, and Neural Networks (NN). The large number of features makes the problem very complex, so we expect the more complex classification technique like SVM and NN to outperform a more straightforward technique like KNN.

We used an F1 score as the measure of accuracy, as it takes into account both precision and recall by counting false negatives and false positives. Specifically, we used an F1 Macro score which takes the arithmetic mean of the F1 scores for each class.

For the KNN model, we first tuned the value of k by examining the cross-validation scores of the testing dataset given different values of k. We found that for both the basic model and the binarized model, a value of $k = 3$ was the highest performing, seen in figure 1. The graph for the binarized model is virtually identical.

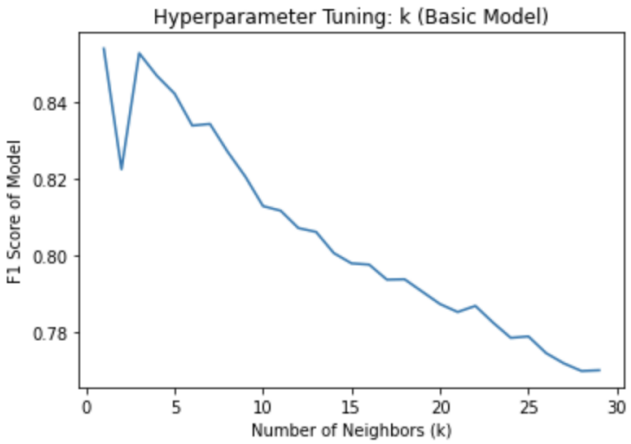

Figure 1: Tuning K for the basic model

Next, for the SGD model, we had to tune two hyperparameters, the L2 regularization weight and the learning weight. We found, using a grid search, that a regularization weight of 0.01, and a learning rate of 0.001 yielded the highest f1 score over the test cross validation.

For the SVM model, we had to tune the parameter C, which is the inverse of the L2 regularization strength. We found that a value of $C = 7$ yielded the highest f1 score. This can be seen in figure 2.

For the random forest model, we had to tune the hyperparameter "n_estimators," which is the total number of trees in the forest. Like the previous tuning, we tested the performance over a range of values, and as can be seen in figure 3, we found that a value of n_estimators = 235 yielded the best score.

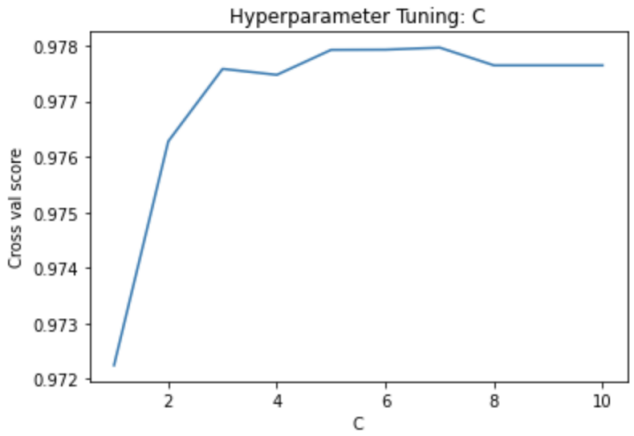

Figure 2: Tuning K for the basic model

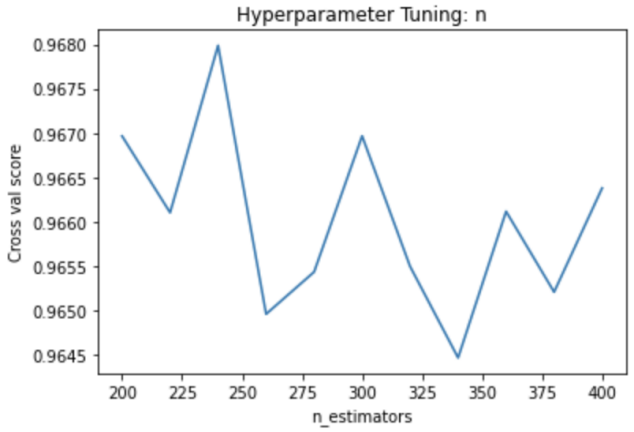

Figure 3: Tuning K for the basic model

Lastly, for the neural network, we tuned two hyperparameters, the regularization weight and the learning rate. This was also done with a grid search. We found that a regularization weight of 0.01 and a learning rate of 0.001 resulted in the highest accuracy score.

5 Results

First, we examined the efficacy of our machine-learning models on the feature-binarized dataset. After we tuned the hyperparameter of every model, we found that the

best-performing model was SVM, with $C = 7.0$ and using the kernel rbf, resulting in a mean cross-validation score of 0.978 (Figure 4).

Binarized Model Results

Model	Cross val score
Neural Net	0.972
SVM	0.978
LinearSVM	0.972
SGDClasssifier	0.972
KNN	0.855
Random Forest	0.968
Baseline Model	0.379

Figure 4: Mean cross-validation scores of various models after 10 folds of cross-validation

Then, we applied the SVM model to the test set, using the f1 score as an evaluation metric and comparing it against the performance of a baseline model that outputs the most common cancer type (mode) of the training labels. We obtained an f1 score of 0.973 for SVM, while the f1 score of the baseline model was 0.379. Evidently, the SVM model significantly outperforms the baseline model. Taking into account our high accuracy scores, we can conclude that our support vector machine model, with feature-binarization, most accurately classifies cancer types using miRNA isoform reads as features.

We then examined the effects of feature-binarization by removing our feature binarization process and instead min-max scaling our data.

Min/Max Scale Model Results

Model	Cross val score
Neural Net	0.969
SVM	0.965
LinearSVM	0.965
SGDClasssifier	0.972
KNN	0.865
Random Forest	0.976
Baseline Model	0.379

Figure 5: Mean cross-validation scores of various models after 10 folds of cross-validation without feature-binarization

Though the general performance of models does not alter much, the most optimal model for the non-feature selected dataset was the random forest classifier with the parameter n estimators set at 100 (Figure 5). We then ran the random forest classifier against the baseline and obtained an f1 score of 0.966, compared to the baseline which was 0.379.

With the conclusion of the evaluation of our models against the test set, we can state that the most optimal model for cancer classification is SVM with C set at 7.0 and the features binarized, as it marginally outperforms the random forest classifier.

6 Broader Impacts

With the rapid technological advancement in machine learning and biology, researchers are now able to apply machine learning models to assist with public health problems such as the classification of cancer types. The more we understand about biomarkers' expression patterns in cancer types, the more capable we are in the ability to accurately diagnosing patients with cancer, which in turn will lead to better treatment and chances of recovery.

However, it is important to take into account the potential errors of these models. It is imperative that we rely not only on machine learning classification in diagnosis. Doctors must also take into account other biological aspects of cancer that machine learning cannot account for. The dangers of misclassification of cancer type are immense; if a patient's cancer type is misclassified, treatment plans may have no or even adverse effects on the patient. Additionally, each case of cancer is unique, which exposes an area of weakness for machine learning models. Misclassification of novel cancers could also lead to disastrous consequences.

In the future, a more comprehensive machine learning model that takes in features beyond just miRNA isoforms may lead to an even more accurate classification of cancer types. Researchers and scientists should continue to strive for 100 percent classification accuracy to better improve diagnosis of cancer and thus survival rates of patients.

7 Conclusion

In our experiment, we found that the support vector machine, using the kernel rbf, was the best model in classifying 6 cancer types using miRNA isoforms. This was surprising because given the number of features, we would normally expect a linear kernel svm to outperform an rbf kernel svm.

In the future, we can experiment with more forms of feature selection beyond binarizing the miRNA isoforms. Hopefully, a model with more features will be able to score higher. Additionally, we could get data on different cancer types to expand the practicality of our model.

8 Contributions

IR and JC both formatted the dataset, IR binarized the variables and coded the NN, JC coded the rest of the models. IR wrote the dataset, exploratory analysis, experiments, and conclusion sections. JC wrote the abstract, intro, results, broader impacts, and conclusion sections. Both authors proofread the paper.

References

Grossman, R. L.; Heath, A. P.; Ferretti, V.; Varmus, H. E.; Lowy, D. R.; Kibbe, W. A.; and Staudt, L. M. 2016. Toward a shared vision for cancer genomic data. *The New England Journal of Medicine* 375(12):1109–1112.

Lopez-Rincon, A.; Mendoza-Maldonado, L.; Martinez-Archundia, M.; Schönhuth, A.; Kraneveld, A. D.; Garssen,

J.; and Tonda, A. 2020. Machine learning-based ensemble recursive feature selection of circulating mirnas for cancer tumor classification. *Cancers* 12(7).

Telonis, A. G.; Magee, R.; Loher, P.; Chervoneva, I.; Londin, E.; and Rigoutsos, I. 2017. Knowledge about the presence or absence of miRNA isoforms (isomiRs) can successfully discriminate amongst 32 TCGA cancer types. *Nucleic Acids Research* 45(6):2973–2985.