STA561： Machine Learning Method in Predicting Glioma Mortality Based on Genetic and Clinical Biomarkers

Author: Baiying Lu, Sumin Lan (Department of Biomedical Engineering)

ABSTRACT

Low grade gliomas are sub-types of cancer that develops in the glial cells of the brain, which are fatal diseases happen world-wildly. In this study, the genetic biomarkers' expression Z scores and clinical data from electronic health record in TCGA-LGG are used as the databases to build machine learning models on. After a missing data imputation and a brief data pre-processing, feature selection methods including partial least square and principal component regression are used to reduce the dimension of the database. Then, classifier including logistic regression, random forest and convolutional neural network are applied to predict the glioma mortality in patient. The result shows that genetic information is persuasive enough in predicting glioma mortality. Clinical data could be helpful only in some of the cases. And CNN model performs the best to make a decent prediction on patient glioma mortality.

BACKGROUND AND DATA

As the data from Figure. 1, patient, especially aged from 50 to 85, has higher probability to have Glioma these days, so it is necessary to analyze this kinds of brain cancer. Therefore, we want to do research on the mortality of this disease.

In this study, TCGA-LGG(Low Grade Glioma) database is used to predict the mortality of glioma. TCGA-LGG was collected by The Cancer Genome Atlas (TCGA), a landmark cancer genomics program launched by the National Cancer Institute and the National Human Genome Research Institute in 2006. In the TCGA-LGG database, two data frames are used in this Machine Learning Method in Predicting Glioma Mortality Based on Genetic and Clinical Biomarker study(MLPG study).

We tend to compare 2 kinds of dataset: genetic dataset only and genetic and clinical dataset. We want to find the importance of gene function when predicting the mortality of patients who have glioma. Due to huge amounts of genetic information in features, we just visualize some gene, which we can see the relationship in the figure 2 heatmap.

Fig 1. Giloma Rates by Gender and Age[1]

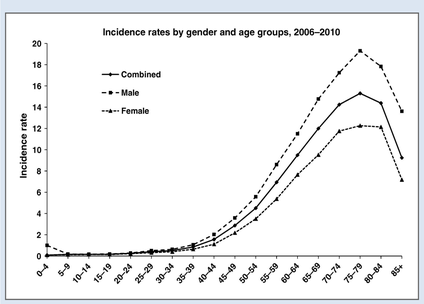
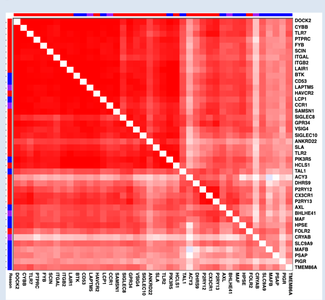


Fig 2. Some Genetic Feature Heatmap[2]



[1] Epidemiologic and Molecular Prognostic Review of Glioblastoma: Jigisha P. Thakkar, Therese A. Dolecek, Craig Horbinski, Quinn T. Ostrom, Donita D. Lightner, Jill S. Barnholtz-Sloan and John L. Villano. Cancer Epidemiol Biomarkers Prev October 1 2014 (23) (10) 1985-1996; DOI: 10.1158/1055-9965.EPI-14-0275

[2] TCGA-LGG(Low Grade Glioma) Database:https://www.cbioportal.org/study/summary?id=lgg_tcga

METHODS AND RESULTS - CLASSIFIERS

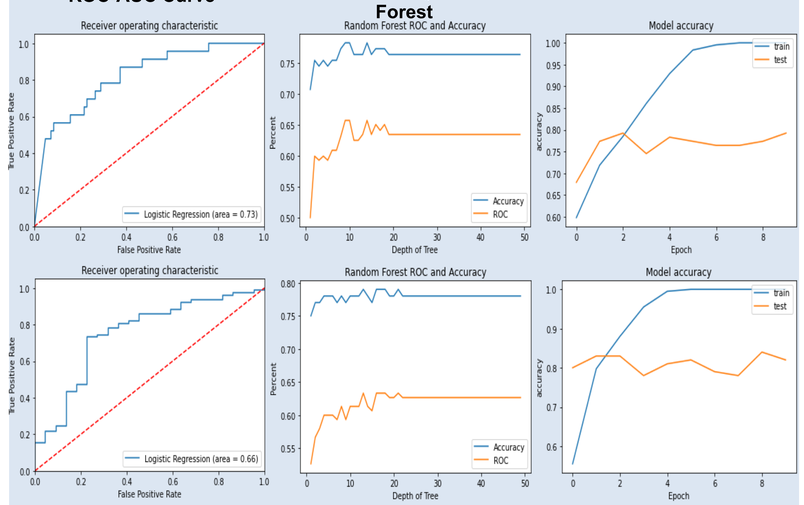
Three classifiers are used in this MLGP study to predict patient mortality. The assumption of this study is that the patient mortality issue is a binary classification problem. Logistic regression is a persuasive method as a binary classifier. Random forest can be applied to data with large quantity of variables. And CNN is a powerful tool modeling on some complicated data , such as image or long sequential data. The application of CNN on genetic data is a new attempt for us.

We try to apply ROC-AUC curve to plot the logistic regression so that we can compare sensitivity and specificity. more intuitively. We visualize the CNN models only with the overall accuracy. These two-row plots illustrate the result of our models, In particular, the first row use the genetic database while the second row use the genetic and clinical data.

Fig 3. Logistic Regression ROC-AUC Curve

Fig 4 Ratio of ROC Score and Overall Accuracy of Radom Forest

Fig 5. Accuracy Of CNN



METHODS AND RESULTS – PLS AND PCR

For genetic database, there are more than 20,532 genetic markers in the database as features. So, feature selection will make the data easier to be interpreted. The hypothesis of this MLGP study is that the feature selection will help the classifier to improve the prediction accuracy and ROC-AUC Score. One special characteristic of genetic database is that the biomarkers may be correlated with each other, which makes LASSO inappropriate to be applied here because LASSO will just randomly pick one from a correlated feature set. So in this MLGP study, PLS and PCR is more suitable.

The following plots demonstrate the principle components of our genetic dataset, it can provide us the process of feature selection intuitively.

Fig 6. Number of Components of PLS

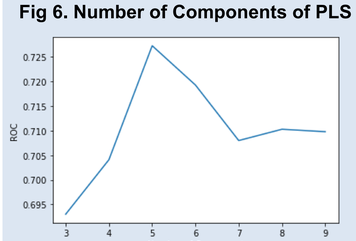
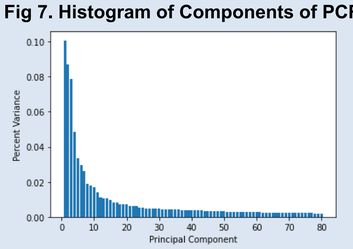


Fig 7. Histogram of Components of PCR



ANALYSIS

When we try to make comparison across all models, it is not hard to seek that CNN model achieves the best accuracy based on either genetic data or clinical data. Just as the reason we have discussed before, CNN model has powerful and advanced self recognition system, so that it can make such decent prediction. The only problem is that it seems like our CNN model is a little bit overfitting. It is mainly because of the parameters. Our epoch is too large to train, we will discuss that in the future work section. With different combination of different modules, 18 different models are generated. The result of each model is reported in the following tables.

Only Gene	PLS	
Classifier	ROC-AUC SCORE	ACCURACY
Logistic Regression	0.631	0.698
Random Forest	0.663	0.792
	Train	Validation
CNN	0.979	0.78

Only Gene	PCR	
Classifier	ROC-AUC SCORE	ACCURACY
Logistic Regression	0.679	0.783
Random Forest	0.612	0.764
	Train	Validation
CNN	1	0.717

Only Gene	No Feature Selection	
Classifier	ROC-AUC SCORE	ACCURACY
Logistic Regression	0.722	0.642
Random Forest	0.588	0.773
	Train	Validation
CNN	1	0.830

Gene and Clinical	PLS	
Classifier	ROC-AUC SCORE	ACCURACY
Logistic Regression	0.592	0.670
Random Forest	0.549	0.710
	Train	Validation
CNN	0.9875	0.79

Gene and Clinical	PCR	
Classifier	ROC-AUC SCORE	ACCURACY
Logistic Regression	0.663	0.760
Random Forest	0.558	0.77
	Train	Validation
CNN	1	0.79

Gene and Clinical	No Feature Selection	
Classifier	ROC-AUC SCORE	ACCURACY
Logistic Regression	0.684	0.750
Random Forest	0.660	0.810
	Train	Validation
CNN	1	0.82

FUTURE WORK

Since overfitting happens in our CNN models, in future, we may consider to optimize the structure of CNN, such as changing the number of layer, changing the kernel size and changing the learning rates. For the logistic regression module, tuning the parameter which is the threshold in this case may increase the performance of the model. So we also plan to use 10-fold cross-validation to find the best threshold.