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

Baiying Lu, Sumin Lan
Department of Biomedical Engineering
Duke University
Durham, NC 27708
baiying.lu@duke.edu, sl597@duke.edu

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.

1 Introduction

1.1 Background

Glioma is a wide category of brain tumor originates from glial cells. According to the World Health Organization (WHO), gliomas can be differentiated into 4 grades (I–IV). Lower-grade gliomas (LGG) ([WHO] grades II and III) are aggressive tumors that occur most commonly in the hemi-cerebrum of adults and include astrocytomas and oligodendrogliomas.[1] Tumors are classified into grades I, II, III or IV based on standards set by the World Health Organization. Regardless of grade, as a glioma tumor grows, it compresses the normal brain tissue, frequently causing disabling or fatal effects.[2]

To gain further understanding of this heterogeneous disease, there have been several scientists conducted integrative genome-wide analysis of 511 LGGs from adults, using bioinformatics analysis.[3] To handle high dimensional data like genetic data, machine learning (ML) algorithms are widely used for creating reliable statistical models for classification and outcome prediction. Therefore, the aim of this study was to apply different kinds of machine learning models to predict the survival of patients with genetic and clinical biomarkers. Our study is only based on lower grade glioma.

1.2 Data Description

In this study, TCGA-LGG(Low Grade Glioma) database is used to predict the mortality of glioma in patients who are suffering or suffered from glioma. The data is an open-source database which can be directly downloaded from the website(https://www.cbioportal.org/study/summary?id=lgg_tcga).

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, molecularly characterized over 20,000 primary cancer and matched normal samples spanning 33 cancer types.

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). The genetic information is mainly from "data_RNA_Seq_v2_mRNA_median_Zscores" which includes the median Z scores of 20,532 genetic biomarkers from 530 patients. And the clinical information is mainly from "data_bcr_clinical_data_patient" which includes the electrical health record of 515 patients from clinics.

2 Methods

The method includes data pre-processing, feature selection methods and machine learning classifiers. Data pre-processing deals with the missing data with dropping and imputation. Feature selection methods mainly describes Partial Least Square(PLS) and Principal Component Regression(PCR). And the classifiers includes Logistic Regression, Random Forest and Convolutional Neural Network(CNN).

2.1 Data Pre-Processing

Our purpose is to predict the mortality of patients, so it is a binary classification. Considering our large amounts of features, there are many variables are in categorical type, we try to convert all these categorical string into columns name in order to enable all these biomarkers as a format of "0" or "1".

Our database comes from 515 patients from different clinics, it seems that there is large amounts of missing data in this dataset. The common skill we need to apply is to do imputation, but we only have 515 observations and except for some numerical features like age, most columns are categorical type, so it is hard to use imputation. Besides, we only have 515 observations, it cannot delete the entire row, so we decide to reduce the variables which have higher than 10% missing data first. After that feature selection, we use mode imputation to replace missing values of categorical variable by the mode of non-missing cases of that variable.

To achieve our goal to apply machine learning models, we create 2 matrices. One of them is based on genetic independent variables, and the other combines genetic and clinical variables together. Then we split them to train dataset, validation dataset(for tuning parameters in each model) and test dataset seperately.

2.2 Feature selection

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(ROC-AUC score will be discussed in the Result and Analysis part). One special characteristic of genetic database is that the some of the geneticmarkers may be correlated with each other, which makes LASSO inappropriate to be applied here because LASSO just randomly picks one from a correlated feature set. So in this MLGP study, PLS and PCR is more suitable.

2.2.1 Principal Component Regression (PCR)

Principal Component Regression (PCR) is a feature selection method based on Principal Component Analysis (PCA). PCA transfers the correlated features into the combination of uncorrelated principal components based on the maximum variance. The loadings from the PCA of training data is extracted as the regression model which is applied on the test data. This process forms some regression equation which maximizes the adjusted R^2 and minimizes the standard error of estimate[4]. In the MLGP study, the first 80 principal components loadings are extracted because 80 principal components can describes 75% variance precisely.

2.2.2 Partial Least Square (PLS)

Partial least squares (PLS) was developed by Herman Wold in 1960ś. As Randall D. Tobias described in "An Introduction to Partial Least Squares Regression," "PLS is a method for constructing predictive models when the factors are many and highly collinear[5]." In this MLGP study, because the genetic biomarkers' correlations exists, the application of PLS is well-reasoned. During the process of PLS model construction, the label is necessary, which means that PLS can be taken as a "supervised" PCR. In addition, it is significant to tune the number of component, so 10-fold cross validation is applied here for tuning the parameter to get the highest accuracy and ROC-AUC score. After standardizing the p predictors, PLS computes the first direction Z_1 by setting each ϕ_{1j} in the following equal to the coefficient from the simple linear regression of Y onto X_j.

$$Z_m = \sum_{j=1}^{p} \phi_{1j} X_j$$

Hence, from the above equation, PLS places the highest weight on the variables that are most strongly related to the response. Subsequent directions are found by taking residuals and then repeating the above prescription[6].

2.3 Machine Learning 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.

2.3.1 Logistic Regression

As we describe above, our problem is a binary classification. Logistic regression as a powerful discriminative method was used to address a binary classification problem. Although logistic regression often performs comparably to competing methods, such as support vector machine and linear discriminant analysis, it is chosen in this problem because of its several advantages.[7] Its theory is based on quantifying the relationships between different types of land use and their drivers, which is specified by

$$log(\frac{p(x)}{1-p(x)}) = \beta_0 + x\beta$$

We use validation dataset to do k-fold cross-validation and our k is equal to 10 to tune the parameter used in logistic regression, for example, inverse of regularization strength or penalty[6].

2.3.2 Random Forest

The tree-based method that can be used for regression and classification problems are called decision tree, which could be used for continuous and categorical outcomes. Decision trees can stratify or segment the predictor space into a number of simple regions and recursively split data into hierarchical subsets

Random forests provide an improvement over bagged decision trees by way of a small tweak that decorrelates the trees, which reduces the variance when the results from all decision trees are averaged . As in bagging, 1000 decision trees are build on bootstrapped training samples.

But when building these decision trees, each time a split in a tree is considered, a random selection of m predictors is chosen as split candidates from the full set of p redictors. The split is allowed to use only one of those m predictors. A fresh selection of m predictors is taken at each split, and typically we choose $m \approx \sqrt{p}$ — that is, the number of predictors considered at each split is approximately equal to the square root of the total number of predictors[6].

In this MLGP study, 10-fold cross validation is applied to tune the depth of the decision trees to get the best prediction accuracy and ROC-AUC code.

2.3.3 Convolutional Neural Network

CNN (Convolutional Neural Networks) models are similar to regular fully connected neural networks. The design of CNN entails many architectural choices to account for number of hidden layers, number of filters, or their size are some, to name a few[8]:

$$\phi = f(\sum_i w_i x_i + b)$$

Figure 1 demonstrates the principle of CNN, which benefit problems like unlabeled data, semi-supervised learning. By adding or shifting several specific physical layer, it can result in different weights. It often being applied to image or video, we also can use it to predict our 'matrix'. Our work is mainly to create CNN model and tune parameters like epoch and learning rate to adjust the model in order to get the best classification.

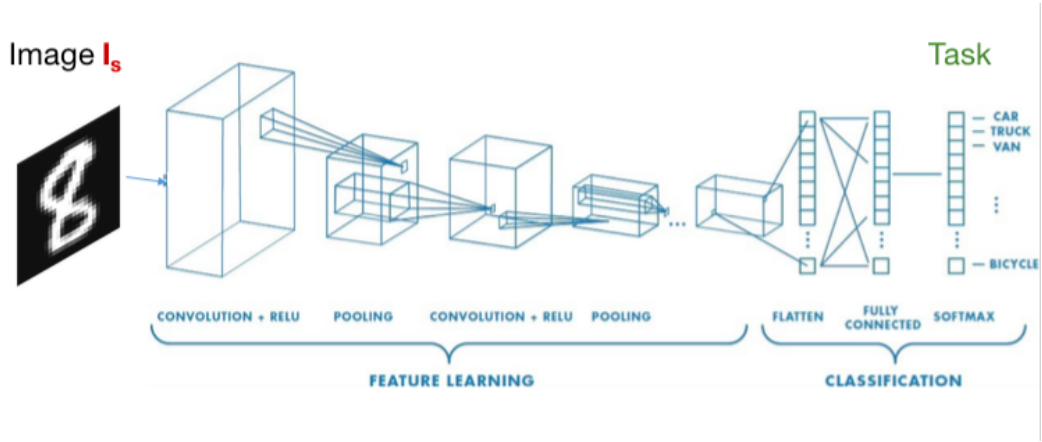


Figure 1: Convolutional Neural Network

3 Result

As we mentioned above, there are three methods in feature selection part: PCR, PLS, without feature selection. And there are three classifiers: logistic regression, random forest and CNN. With different combination of different modules, 18 different models are generated. The result of each model is reported in the following tables.

3.1 Assessment Calibration

For the two classifier, logistic regression and random forest, two model assessment calibration are taken into consideration. The first is ROC-AUC score, the second is accuracy.

Two metrics could be defined based on the confusion matrix, which includes four parts. The first is True Positive (TP), which means that false positive occurs when the true value is negative, but the predicted value is positive. False Positive (FP) means a false positive occurs when the true value is negative, but the predicted value is positive. False Negative (FN) means a false negative occurs when the true value is positive, but the predicted value is negative. True Negative (TN) means A true negative occurs when the true value is negative, and the predicted value is also negative.

Accuracy, using the terms defined above, is equivalent to:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

ROC-AUC score is the score value of area under the ROC curve. Receiver Operating Characteristic (ROC) is the result of Sensitivity over 1-Specificity. Area Under the Curve (AUC) is the area under the ROC curve.

$$ROC = \frac{Sensitivity}{1 - Specificity}$$

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

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

Table 1. Different Assessment Calibrations of Different Models against Different Databases

4 Analysis

4.1 Analysis of Logistic Regression

Table 1 reports individual overall accuracy and ROC-AUC score for each classifier. We can find that the ROC-AUC score and overall accuracy are all fluctuated between 0.65 to 0.75. In particular, the model with PLS has a lowest ROC score based on both database. Different from CNN, PCR is more useful when applying to the logistic model than PLS does. When dealing with the analysis of no feature selection, only genetic database has quite different value between ROC score and general accuracy, which means ROC score is high while accuracy is low. This represents for our logistic classifier currently does a bad job on this dataset, we may have to find suitable threshold, like the value in genetic and clinical database. In general, which is surprisingly similar to CNN, no feature selection method accuracy seems quite decent to predict the mortality of patient.

4.2 Random Forest

In general, the model accuracy based on random forest is more than 70, while the ROC-AUC curve is no more that 70, which might because the database itself is an imbalanced database, so the models cannot perform so well as we assumed before.

Based on the ROC-AUC score, random forest works the best when PLS is the feature selection method only with gene information. For random forest trained only with genetic dataframe, PLS always outperform PCR and no feature selection. While with clinical data, PLS still does a better job than PCR, but random forest can perform the best without feature selection. This phenomenon means that there might be some features in the clinical data is principal. In conclusion, Without clinical data, random forest can perform well with feature selection in which PLS outperforms PCR. While with clinical data, random forest can perform well without feature selection. In addition, with feature selection, random forest can do a better job only with gene information than with clinical data.

4.3 Analysis of Convolutional Neural Network

As we can see from the Table.2, in general, our CNN model is a little bit overfitting through all condition. Since all the train accuracy keep increasing while test accuracy tend to be constant as epoch growing. When comparing Clinical and genetic biomarkers to genetic biomarkers only, we can find that in each row, the tendency of the training and testing accuracy is quite similar. The only difference is that the database with clinical information has a higher accuracy than the original one. Comparing the accuracy in each column, it demonstrates the influence of feature selection methods. Accordingly, PLS and PCR are convenient for data with highly-correlated predictors. The number of PCs used in PLS is generally chosen by cross-validation. Predictors and the outcome variables should be generally standardized, to make the variables comparable. So it seems that in this CNN model, the PCR and PLS method are more difficult to interpret compared with the original one, because they do not perform any kind of variable selection or even directly produce regression coefficient estimates. Besides, because of powerful and advanced function of Convolutional Neural Network, it can adjust parameters and feature importance by itself. However, the when it comes to the 2 feature selection accuracy, due to PLS can use a dimension reduction strategy that is supervised by the outcome, PLS performs better than PCR.

4.4 Discussion about different models

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.

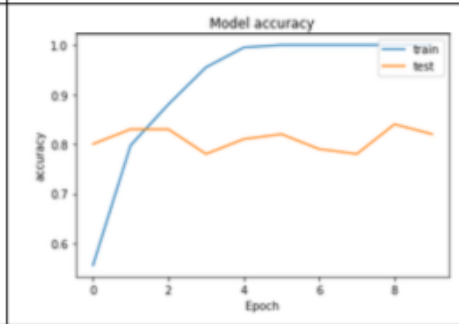
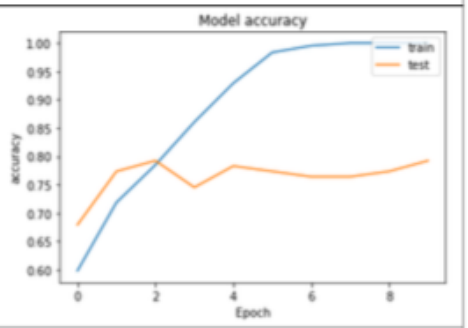
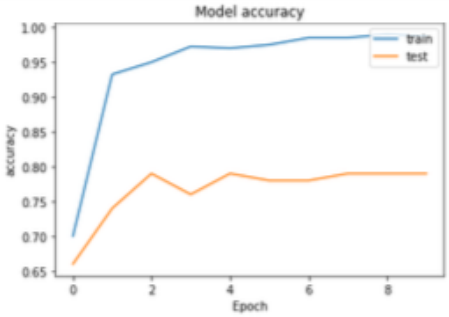
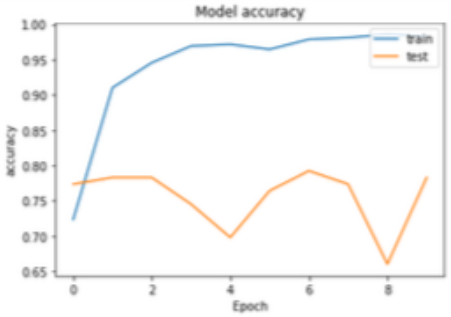
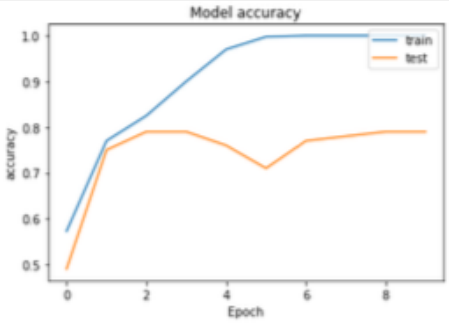
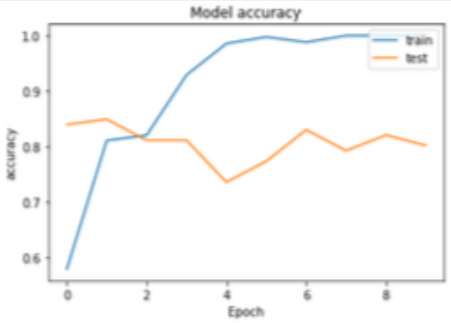
	Clinical and Genetic Biomarkers	Only Genetic Biomarkers
Train and test accuracy without PLS		
Train and test accuracy with PLS		
Train and test accuracy with PCR		

Table 2. CNN Model Accuracy Based On Different Biomarkers And Feature Selection

5 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 rate.

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.

Acknowledgments

We would like to express our special thanks of gratitude to the professor Dr. Sayan Mukherjee as well as our TAs who gave us a lot of the theory support and coding logic help which enable us to finish this project on the topic Machine Learning Method in Predicting Glioma Mortality Based on Genetic and Clinical Biomarkers, which also helped me in doing a lot of Research and we came to know about so many new things. We are really thankful to them.

References

[1] Louis DN, Perry A, Reifenberger G, et al. The 2016 World Health Organization Classification of Tumors of the Central Nervous System: A summary. Acta Neuropathol. 2016;131:803–20.

[2] American Cancer Society. Brain and Spinal Cord Tumors in Adults. Available at: <http://www.cancer.org/Cancer/BrainCNSTumorsinAdults/DetailedGuide/brain-and-spinal-cord-tumors-in-adults-what-are-brain-spinal-tumors> [Accessed June 2011].

[3] Deng, Teng et al. "Use of Genome-Scale Integrated Analysis to Identify Key Genes and Potential Molecular Mechanisms in Recurrence of Lower-Grade Brain Glioma." Medical science monitor : international medical journal of experimental and clinical research vol. 25 3716-3727. 19 May. 2019, doi:10.12659/MSM.913602

[4] R.X. Liu, J. Kuang, Q. Gong, X.L. Hou, Principal component regression analysis with spss, Computer Methods and Programs in Biomedicine, Volume 71, Issue 2, 2003, Pages 141-147

[5] Randall D. Tobias. "An Introduction to Partial Least Squares Regression" 1995. PDF file.

[6] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. An Introduction to Statistical Learning : with Applications in R. New York :Springer, 2013.

[7] Y. Zhu, T.L. Tan, W.K. Cheang Penalized logistic regression for classification and feature selection with its application to detection of two official species of Ganoderma Chemometr. Intell. Lab. Syst., 171 (2017), pp. 55-64

[8] KDEEP: Protein–Ligand Absolute Binding Affinity Prediction via 3D-Convolutional Neural Networks José Jiménez, Miha Škalič, Gerard Martínez-Rosell, and Gianni De Fabritiis Journal of Chemical Information and Modeling 2018 58 (2), 287-296 DOI: 10.1021/acs.jcim.7b00650