Predicting Pathology Features using Gene Expression Data in Glioblastoma Patients

Nishant Subramani

Introduction:

Glioblastoma multiforme (GBM) is the most common, malignant and aggressive primary brain tumor that affects humans. Survival rates are low--the median survival rate without treatment is 4.5 months, but with standard-of-care and chemotherapy, this rises to 15 months (Johnson 2011, Van Meir 2010). Gene sequencing technologies have developed greatly since the human genome project and patients can get genomes sequenced for relatively cheap. This has led to a rise in gene expression analysis and a need for identifying genetic bases for diseases.

Objective:

To predict pathology features in patients with Glioblastoma ascertained from imaging using only patient's clustered gene expression data.

Data:

Data was acquired from Professor Olivier Gevaert at Stanford University. Dataset includes three files, one pathology image data file, and two gene expression files. The pathology image data had 210 outcomes and 224 patients, while the two gene expression files included the same 450 patients but had 100 features and 200 features respectively. Each feature in the gene expression file corresponded to a cluster's expression level which was clustered using an algorithm similar to k-means. These datasets were combined and two datasets with common patients were created, one with 200 features and 6 outcomes, and another with 100 features and 6 outcomes. Both of these datasets have the same 200 patients. All analysis was done in R v3.20. Separate analyses were done for each of the outcomes in each of the new datasets.

Regression Methodology:

All features and outcomes are continuous valued and thus regression techniques are the most appropriate. Three regression techniques were employed: Linear Regression with Univariate Spearman Correlation, Random Forest of Regression CART Trees, and Regression CART Tree. The Linear Regression with Univariate Spearman Correlation used hand-picked p-value thresholds which tried to remove most of the predictors in each dataset to eliminate uncorrelated variables and prevent over-fitting. The Spearman Correlation thresholds were held constant across datasets for each outcome variable and the specific values for each outcome are shown in Table 1. The random forest used the default R settings with 2000 trees being built and the CART tree used the default R settings as well. 10-Fold cross-validation was employed for each of the algorithms due to the small sample size which prevented the splitting of a training, test, and validation set.

Classification Methodology:

For classification to be appropriate in this domain, each outcome had to be discretized. For each outcome, in each fold of cross-validation, a threshold value was chosen by taking the 75th percentile element in the training set. Every value above this was termed risk and given a value of 1, and every value below this was termed a control and given a value of 0. Using this discretization, 4 classification algorithms were employed: Naïve Bayes, Neural Network, Random Forest of Classification CART Trees, and Classification CART Tree. For each model, the default R settings were used, except that the neural network was initialized to have 4 hidden layers and the Random Forest to have 2000 trees. 10-Fold cross-validation was employed again for each algorithm.

Results and Discussion:

Mean squared error (MSE) is the metric of choice for the evaluation of the regression models. This takes the sum of the squared differences between each test set example and predicted value and takes its average. The results of the regression experiments for each of the outcomes on the 100 feature dataset are in Table 2, while the results on the 200 feature dataset are in Table 3. In Tables 2 and 3, random forest and linear regression perform similarly, while the CART tree has a two times higher MSE. This is due to the fact that both the linear regression model and the

random forest do a good job on noisy datasets which have non-complex relationships. Since the random forest is an ensemble method of CART trees, it's expected that if the relationship is relatively non-complex, random forest will perform better than a single CART tree. Area under the receiver operating characteristic (ROC) curve (AUC) is the metric of choice for the evaluation of the classification models. This takes the sensitivity and specificity into consideration and evaluates the models for overall performance, not simply a percent accuracy metric. The results of the classification experiments for each of the outcomes on the 100 feature dataset are in Table 4, while the results on the 200 feature dataset are in Table 5. In Tables 4 and 5, Naïve Bayes, Random Forest, and CART trees all perform very similarly poorly, while the neural network performed very well with an AUC of around 0.85. This shows that the discretization of the outcome variables provides some very complex non-disjunctive relationship between gene expression data and each of the pathology outcomes.

Futures:

Gaining domain knowledge to build more informative models with causal possibilities is one major extension. Another is to identify and utilize raw gene expression data, and use other feature selection methods to garner a better representation of gene expression rather than this specific clustering methodology provided.

Tables:

Table 1: Spearman Correlation Thresholds for each Outcome

Outcome Variable	Spearman Correlation P-Value Threshold
Cellularity	0.0001
Cell Voronoi Area	0.00001
Cytoplasm Background Intensity	0.01
Edge Length	0.0001
Nucleus Area	0.01
Nucleus Background Intensity	0.01

Table 2: Regression Experiment Results for 100 Features

	Cellularity	Cell Voronoi Area	Cytoplasm Background Intensity	Edge Length	Nucleus Area	Nucleus Background Intensity
Lin Reg MSE	1.11235	0.92206	1.10739	0.94126	1.03336	1.06527
Random Forest MSE	1.03765	0.91624	1.10384	0.91109	0.96657	1.08113
Cart MSE	2.04982	1.75909	2.20726	1.97758	2.01612	1.91001

Table 3: Regression Experiment Results for 200 Features

	Cellularity	Cell Voronoi Area	Cytoplasm Background Intensity	Edge Length	Nucleus Area	Nucleus Background Intensity
Lin Reg MSE	1.04567	1.00188	1.30017	1.07246	1.11699	1.14255
Random Forest MSE	1.05734	0.87985	1.13375	0.89069	0.97212	1.08916
Cart MSE	2.03411	1.92832	1.90432	1.86973	1.82195	1.82973

Table 4: Classification Experiment Results for 100 Features

	Cellularity	Cell Voronoi Area	Cytoplasm Background Intensity	Edge Length	Nucleus Area	Nucleus Background Intensity
Naïve Bayes AUC	0.59	0.6066	0.5166	0.6130	0.6188	0.5525
Neural Network AUC	0.85	0.91	0.9407	0.8751	0.8568	0.8947
Random Forest AUC	0.58	0.5733	0.5193	0.5753	0.5197	0.5261
Cart AUC	0.5967	0.5667	0.5358	0.5780	0.5124	0.5138

Table 5: Classification Experiment Results for 200 Features

	Cellularity	Cell Voronoi Area	Cytoplasm Background Intensity	Edge Length	Nucleus Area	Nucleus Background Intensity
Naïve Bayes AUC	0.61	0.6233	0.5430	0.6199	0.6188	0.5886
Neural Network AUC	0.8067	0.8366	0.7662	0.7705	0.7023	0.8152
Random Forest AUC	0.58	0.5333	0.5064	0.5516	0.5336	0.5031
Cart AUC	0.55	0.5467	0.5426	0.5413	0.5354	0.5202

References:

A. Liaw and M. Wiener (2002). Classification and Regression by randomForest. R News 2(3), 18--22.

Andreas Alfons (2012). cvTools: Cross-validation tools for regression models. R package version 0.3.2.

Arne Henningsen and Ott Toomet (2013). miscTools: Miscellaneous Tools and Utilities. R package version 0.6-16.

H. Wickham. ggplot2: elegant graphics for data analysis. Springer New York, 2009.

Hadley Wickham (2015). stringr: Simple, Consistent Wrappers for Common String Operations. R package version 1.0.0.

Henrik Bengtsson (2015). R.matlab: Read and Write MAT Files and Call MATLAB from Within R. R package version 3.2.0.

Jeremy VanDerWal, Lorena Falconi, Stephanie Januchowski, Luke Shoo and Collin Storlie (2014). SDMTools: Species Distribution Modelling Tools: Tools for processing data associated with species distribution modelling exercises. R package version 1.1-221.

Johnson, Derek R.; O'Neill, Brian Patrick (2011). "Glioblastoma survival in the United States before and during the temozolomide era". *Journal of Neuro-Oncology* 107 (2): 359–64.

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Van Meir, E. G.; Hadjipanayis, C. G.; Norden, A. D.; Shu, H. K.; Wen, P. Y.; Olson, J. J. (2010). "Exciting New Advances in Neuro-Oncology: The Avenue to a Cure for Malignant Glioma". *CA: A Cancer Journal for Clinicians* 60 (3): 166–93.

Venables, W. N. & Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0

Weihs, C., Ligges, U., Luebke, K. and Raabe, N. (2005). klaR Analyzing German Business Cycles. In Baier, D., Decker, R. and Schmidt-Thieme, L. (eds.). Data Analysis and Decision Support, 335-343, Springer-Verlag, Berlin.