Predicting Cancer Stage Using miRNA

Anurag Bhattrai, Nathan Byers, Xi Rao, Ed Simpson

Tuesday, April 28, 2015

Introduction

Kidney Renal Clear Cell Carcinoma (KIRC) is the most common form of kidney cancer in humans and is responsible for up to 95% of kidney cancer. As with all cancers, tumor stage can be the most important factor in predicting prognosis and survivability. There are 5 primary stages of disease in cancer, the definitions of which vary slightly based on the type of cancer being described. Stage 0 is reserved for the origination of a tumor and is largely not detected in practice. For KIRC, Stage I is defined as a tumor of 7cm diameter or smaller and restrained to the kidney. Stage II is a tumor larger than 7cm and restrained to the kidney. Stage III involves metastases to a nearby lymph node but not distant organs and may include metastases to fatty tissue or large veins leading from the kidney. Stage IV is a tumor that has spread through the fatty tissue surrounding the kidney, involvement of more than one lymph node close to the kidney and any lymph node not near the kidney as well as distant metastases to other organs. 5 year disease-specific survival rates vary greatly in KIRC, with Stage I being 95%, Stage II at 88%, Stage III at 59% and Stage IV at 20% (Sachdeva 2014). Early stage diagnosis greatly increases the survival rate, although it is difficult to make an early diagnosis due to the molecular complexity and divergent clinical behavior of KIRC patients. MicroRNAs (miRNAs) are small noncoding RNAs that regulate gene expression and influence cell state and phenotype. Overwhelming evidence indicates a causal role of miRNAs in the onset and maintenance of cancer (Croce 2009). Our study aims to develop classification models to distinguish early stage and late stage of KIRC base on miRNA expression profiles.

Data

The dataset for this study was downloaded from The Cancer Genome Atlas (TCGA) data protal (https://tcga-data.nci.nih.gov/tcga/tcgaHome2.jsp). The Level 3 miRNASeq data for KIRC on Illumina GA and HiSeq miRNA sequencing platform consists of expression of 1046 miRNAs for 617 samples, within which 536 are tumor samples. The clinical information for KIRC was obtained from the "clinical Biotab" section of the data matrix, including the information of 531 patients. Through matching BCR (Biospecimen Core Resource) IDs, the miRNA expression of these 531 patients were retrieved from the miRNASeq dataset.

We generated a miRNA expression data matrix in Comma Separated Values (CSV) file format from the TCGA data in `R`, with 1046 miRNAs as column labels and 531 patients BCR IDs as row labels. The normalized counts, termed reads per million miRNA mapped, were used as an estimate for miRNA expression. The miRNAs (total 201) with no expression across all the 531 patients were removed from the data matrix, leaving 845 miRNAs for the following analysis.

The "ajcc pathologic tumor stage" from TCGA clinical data was used as tumor stage. Within 531 KIRC patients, 266 are at Stage I, 57 are at Stage II, 127 are at Stage III, and 81 are at Stage IV. We marked class label of "Early Stage" for patients with clinical tumor stage I & II (total 323), and class label of "Late Stage" for tumor Stage III and IV (total 208), for the following analysis.

This study dataset was randomly stratified and split into two groups: 80% as the training dataset and 20% as the independent testing dataset.

Methods

Feature selection

To reduce VC dimensions, we first performed a pre-selection step on the training set to keep the top significant features correlated with tumor stage. We tried two methods. In the first method (method 1, and subsequently labeld as "PRE1" in figures and tables), we filtered out miRNAs that do not have at least 20% of the sample expression value greater than or equal to 100 and do not have coefficient of variation (sd/mean) between 0.7 and 10, by using genefilter from Bioconductor (http://bioconductor.org/biocLite.R). 45 miRNAs were left after filtering. In the second method (method 2, and subsequently labeled "PRE2" in figures and tables), we first calculated the correlation of miRNA expression with tumor stage (early or late), and then narrowed the features down by choosing a p-value threshold of 0.1. We then used principal component analysis to transform the data into linear combinations of the miRNA expression data (account for 95% of the variation in the data). 117 miRNAs were selected using this second method.

Classification algorithms

We applied three algorithms to the data for predicting the binary outcome of tumor stage: Random Forest, Support Vector Machine (SVM), and Naive Bayes. Random Forest is randomly constructed ensemble of independent decision trees. SVM are a set of supervised learning methods and we used Radial Basis Function Kernel. Naive Bayes algorithm works on the assumption that all the features are statistically independent and is based on Bayes theorem.

We used the `R` package `caret` to train all the models. The `caret` package (which stands for "classification and regression training"") is a set of functions that attempt to streamline the model training process for complex regression and classification problems. We used the `train()` function for training the models. Different parameters were tuned according to each algorithm to get the optimized training models. For random forest (`method='rf'`), the `mtry` parameter was tuned; for SVM (`method='svmRadial'`), the sigma and cost values were tuned; for Naive Bayes (`method='nb'`), fL and usekernel were tuned.

Training models

The three supervised machine learning algorithms were trained on the training dataset and further validated by 10-fold cross-validation. The training models generated were compared based on the accuracy and AUC.

Testing models

The performance of best-trained and cross-validated models were further evaluated on the testing data. Accuracy and AUC were compared.

Results

Random Forest

Table 1 shows the accuracy, sensitivity, and specificity for the models that were trained using the Random Forest algorithm.

The accuracy was very similar (if not identical) among all four models. Sensitivity and specificity were also very close among the four models. Specificity was much higher than sensitivity, which means that the model was much more accurate predicting an early stage diagnosis. Figure 1 shows the ROC curves for each model.

	Preprocess	Split	Accuracy	Sensitivity	Specificity
1	PRE1	10CV	0.75	0.51	0.89
2	PRE1	5CV	0.74	0.49	0.89
3	PRE2	10CV	0.75	0.49	0.91
4	PRE2	5CV	0.75	0.54	0.88

Table 1: Random Forest Results

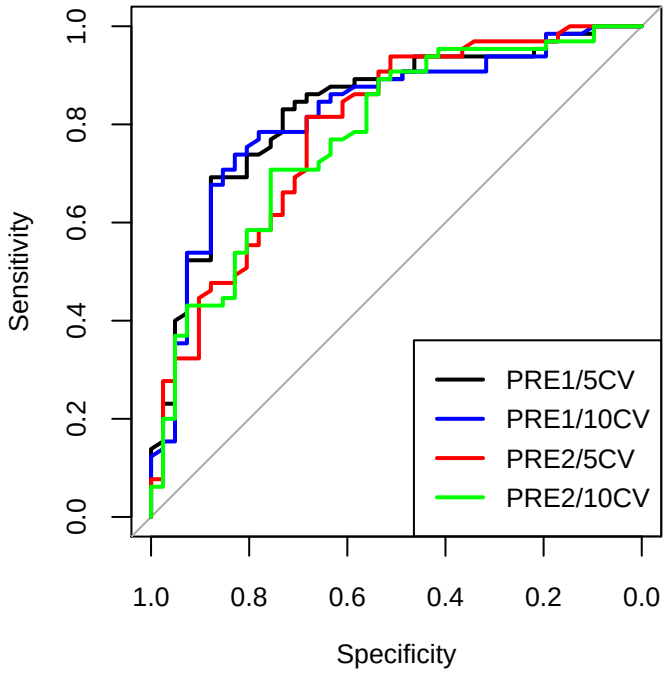

Figure 1: Random Forest ROC curves

Support Vector Machine

Table 2 shows the accuracy, sensitivity, and specificity for the models that were trained using the Support Vector Machine algorithm.

	Preprocess	Split	Accuracy	Sensitivity	Specificity
1	PRE1	10CV	0.69	0.39	0.88
2	PRE1	5CV	0.73	0.49	0.88
3	PRE2	10CV	0.68	0.41	0.85
4	PRE2	5CV	0.70	0.39	0.89

Table 2: Support Vector Machine Results

The accuracy, sensitivity, and specificity were also very similar among all four models. As with the Random Forest models, specificity was much higher than sensitivity. Figure 2 shows the ROC curves for each model.

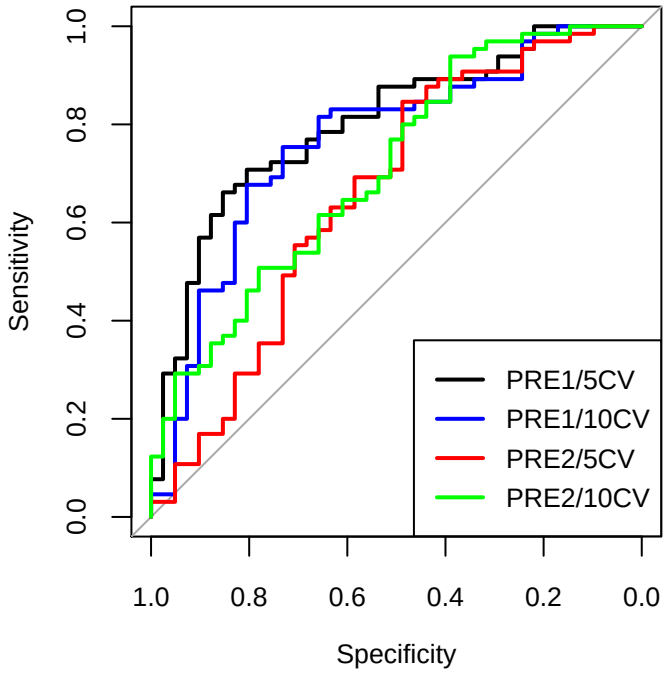

Figure 2: Support Vector Machine ROC curves

Naive Bayes

Table 3 shows the accuracy, sensitivity, and specificity for the models that were trained using the Naive Bayes algorithm.

	Preprocess	Split	Accuracy	Sensitivity	Specificity
1	PRE1	10CV	0.66	0.27	0.91
2	PRE1	5CV	0.64	0.49	0.74
3	PRE2	10CV	0.62	0.37	0.78
4	PRE2	5CV	0.62	0.44	0.74

Table 3: Naive Bayes Results

As with the other algorithms, the accuracy, sensitivity, and specificity were very similar among all four models, with specificity being much higher than sensitivity. Figure 2 shows the ROC curves for each model.

Comparing Algorithms

Figure 4 shows the accuracy results for all four algorithms. Random Forest had the best performance among the algorithms. The algorithms also appear to make the bigest difference in performance. Preprocessing appears to have made little difference for the Random forest algorithm accuracy, whereas it had some impact on the other two alogrithms. The first preprocessing method, which filtered out miRNAs that do not have at least 20% of the sample expression value greater than or equal to 100 and do not have coefficient of variation (sd/mean) between 0.7 and 10, appears to have yielded slightly better accuracy across the algorithms. The two data splitting methods didn't have a consistent effect on accuracy across the algorithms.

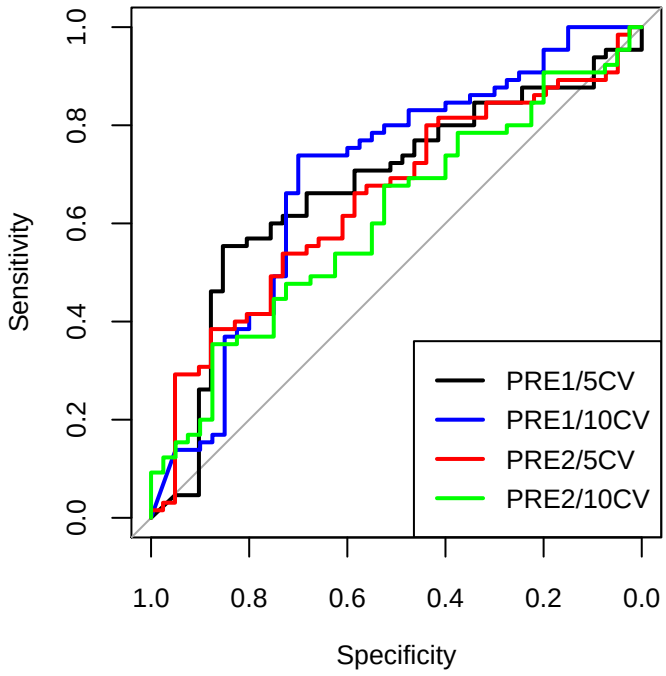

Figure 3: Naive Bayes ROC curves

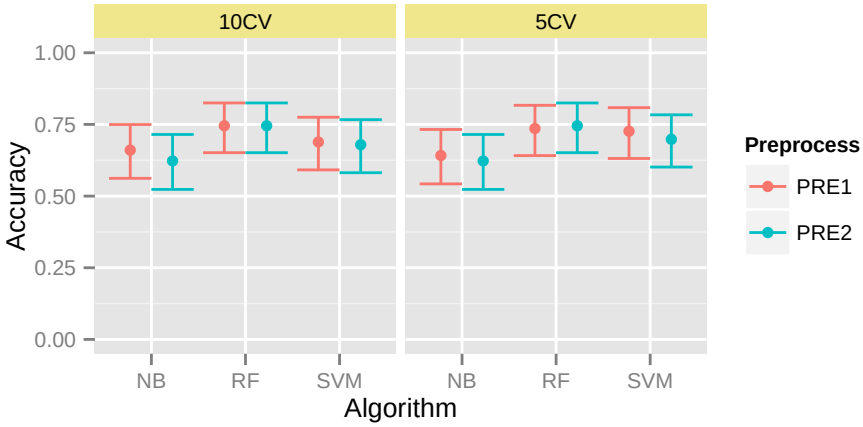

Figure 4: Accuracy for all alogrithms

Figure 5 shows the sensisitvity and specificity for all models for the three alogrithms. As already noted, sensitivity was poor accross all algorithms. Specificity was high for most models.

Considering just specificity, Naive Bayes had the poorest performance among the algorithms. Preprocessing and data splitting appears to have made little difference across the algorithms, except for Naive Bayes.

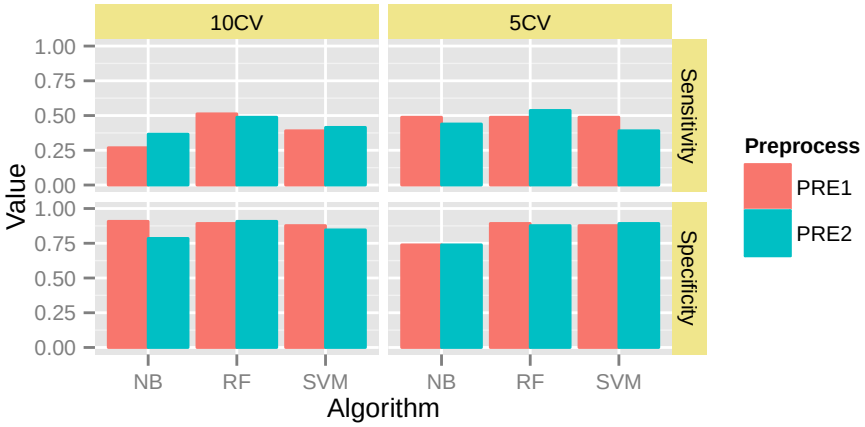

Figure 5: Sensitivity and specificity for all algorithms

We compare the AUC for each model in Figure 6. According to this measurement of performance, the Random Forest alogorithm had the best results. Preprocessing seems to have made a consistent impact on the AUC across the algorithms. Again, the first preprocessing method yielded better performance, and it had a bigger impact than data splitting.

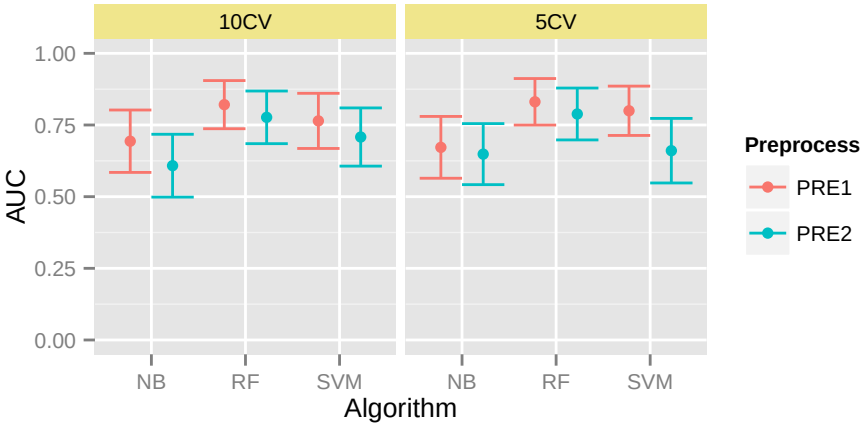

Figure 6: AUC for all algorithms

Discussion

References

Croce, Carlo M. 2009. "Causes and Consequences of MicroRNA Dysregulation in Cancer." *Nature Reviews Genetics* 10 (10). Nature Publishing Group: 704–14.

Sachdeva, Kush. 2014. "Renal Cell Carcinoma." http://emedicine.medscape.com/article/281340-overview.