Northwestern University Feinberg School of Medicine

Evaluation of Data Discretization Methods for Cross Platform Transfer of Gene-expression based Tumor Subtyping Classifier

Segun Jung, PhD, Yingtao Bi, PhD, and Ramana V. Davuluri, PhD

Division of Health and Biomedical Informatics, Department of Preventive Medicine, Northwestern University Feinberg School of Medicine

Background

- Discretization is a technique that transforms continuous values of feature variables to discrete ones. In this study, we investigate different data discretization methods combined with popular machine learning algorithms to derive platform-independent and accurate multi-gene signatures for molecular classification of tumor samples.

Research Objective

- Molecular stratification of tumors is essential for developing personalized therapies. While patient stratification strategies have been successful, computational methods to accurately translate gene signatures derived from high-throughput platforms to clinically adaptable low-dimensional platforms are currently lacking. We perform comparative evaluation of different data discretization methods to derive accurate and platform-independent gene signatures for molecular classification of tumors.

Methods

- Isoform-level gene expression estimates, fold-change expression (tumor over normal) values, and molecular subtype information were obtained for 342 and 155 glioblastoma multiforme (GBM) samples profiled by Affymetrix exon-arrays and RNA-seq, respectively. A subset of 76 GBM samples has expression profiles from both RNA-seq and exon-array platforms. The subtype information or class labels—neural (N), proneural (PN), mesenchymal (M) and classical (CL)—were obtained from our recently published study[1].
- For variable selection, we employed three algorithms, coefficient of variation (CV), support vector machine based (SVM-RFE) and random forest based (RF_based_FS) methods.
- Utilizing the features selected by these two methods, we transformed the fold change data to discretized values by using equal width binning (Equal-W) and equal frequency binning (Equal-F), and Kmeans.
- We use four classification methods—SVM, RF, NB (naïve Bayes), and PAM (prediction analysis of microarray)—to predict a subtype of the GBM samples. Prediction results for independent platform transition were reported based on the 76 common samples.

Results

Table 1. Comparison of classification methods both trained (257 samples) and tested (85 samples) on exon-array data. We report the best accuracy using all available features.

Feature selection	CV (%)				SVM-RFE (%)				varSelRF (%)			
Classifier	FC	Equal-W	Equal-F	Kmeans	FC	Equal-W	Equal-F	Kmeans	FC	Equal-W	Equal-F	Kmeans
SVM	97.6 (1000)	97.6 (600)	97.6 (1300)	96.5 (600)	97.6 (1000)	96.5 (1100)	97.6 (1000)	97.6 (1000)	96.5 (180)	92.9 (80)	98.8 (320)	100 (580)
RF	96.5 (1500)	95.3 (1100)	92.9 (900)	95.3 (600)	97.6 (800)	95.3 (1000)	95.3 (400)	95.3 (400)	95.3 (240)	94.1 (160)	92.9 (420)	95.3 (440)
NB	94.1 (700)	94.1 (900)	92.9 (600)	92.9 (700)	90.6 (2000)	88.2 (2000)	89.4 (2000)	87.1 (1400)	95.3 (280)	95.3 (200)	95.3 (380)	95.3 (720)
PAM	94.1 (1500)	94.1 (1200)	94.1 (1200)	92.9 (1100)	94.1 (2000)	89.4 (1700)	89.4 (1700)	90.6 (2000)	94.1 (620)	94.1 (200)	94.1 (740)	94.1 (640)

Department of Preventive Medicine

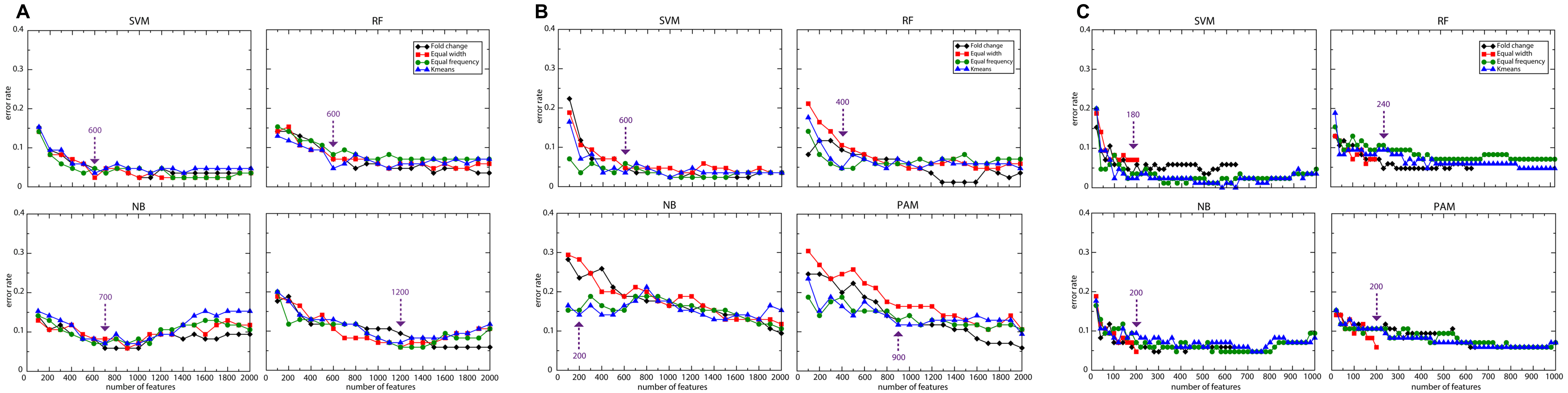

Figure 1. Accuracy of classifiers using features ranked by CV **(A),** SVM-RFE **(B)**, and RF_based_FS **(C).** Independent exon-array dataset of 257 and 85 samples are used for training and testing, respectively. Arrows indicate a cutoff value of features used to test platform transition on RNA-seq dataset

Table 2. Comparison of classification methods trained on exon-array (342 samples) and tested on RNA-seq (155 samples). We report the best accuracy using all available features.

Feature selection	CV	SVM-RFE	RF_based_FS
SVM	97.4 (400)	98.7 (600)	94.7(180)
RF	94.7 (200)	90.8 (300)	93.4 (120)
NB	92.1 (500)	84.2 (200)	92.1 (100)
PAM	93.4 (1100)	89.5 (400)	90.8 (100)

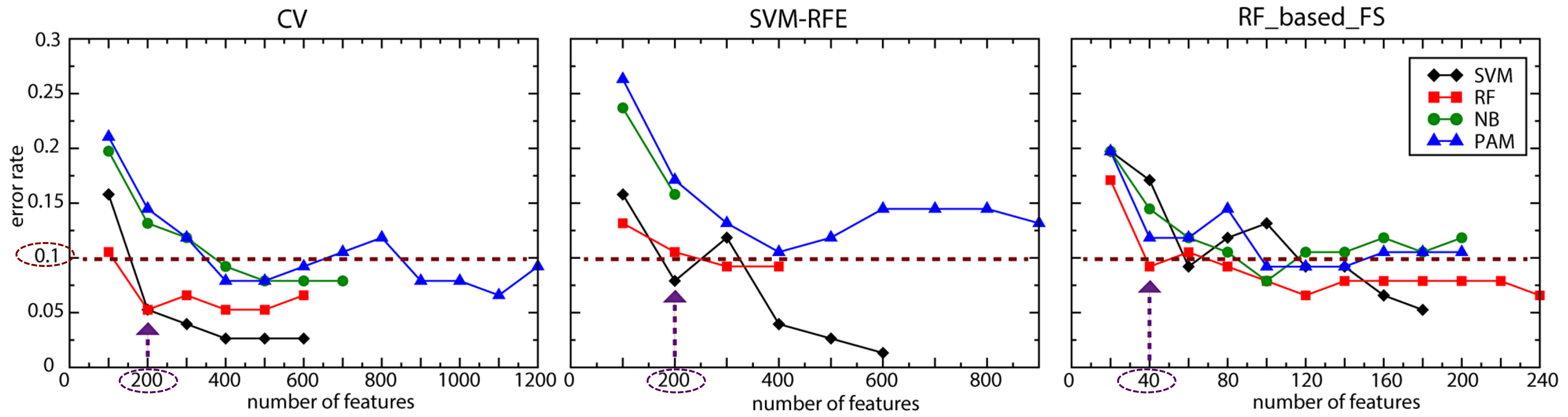

Figure 2. Performance of classifiers trained on exon-array data (342 samples) and tested on RNA-seq data (155 samples) coupled with Equal-F discretization. Features were ranked by CV, SVM-RFE, and RF_based_FS. The dotted brown line marks the 90% accuracy and the purple arrows indicate the minimum number of features to achieve higher than 90% accuracy. The maximum number of features were adopted from the cutoff values shown in Figure 1.

Limitations

- We demonstrated our approach using the GBM dataset. Although applying to diverse dataset (e.g., ovarian cancer) is desirable, cancer patient (class) labels of RNA-seq dataset are currently limited.

Conclusions

- This is the first attempt to tackle multi-class classification problems in cancer genomics using data from different platforms.
- Our experiment showed that SVM and RF coupled with the equal frequency binning discretization and RF based feature selection methods led to robust and accurate classification model when applied to the GBM tumor patients.
- Our approach is generally applicable to other cancer types for molecular classification and identification of subgroups.
- The derived classification models with manageable number of genes/isoforms can be translated to a low-dimensional platform such as high throughput RT-qPCR assays for clinical application. This will provide quantitative and reproducible stratification of cancer patients with prognostic significance, the potential to improve precision therapy and the selection of drugs with subtype-specific efficacy.

NORTHWESTERN UNIVERSITY
FEINBERG
SCHOOL OF MEDICINE

1. Pal et al. *Isoform-level gene signature improves prognostic stratification and accurately classifies glioblastoma subtypes* . Nucleic Acids Res. (2014)