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

Segun Jung, Yingtao Bi, and Ramana V. Davuluri
Division of Health and Biomedical Informatics, Department of Preventive Medicine
Northwestern University Feinberg School of Medicine, Chicago, IL
segun.jung@northwestern.edu, yingtao.bi@northwestern.edu, ramana.davuluri@northwester.edu

Abstract— Molecular stratification of tumors is essential for developing personalized therapies. While patient stratification strategies have been successful, computational methods to accurately translate and integrate gene signatures across different high-throughput platforms (e.g., microarray, RNA-seq) are currently lacking. We performed comparative evaluation of different data discretization and feature selection methods combined with state-of-the-art machine learning algorithms to derive platform-independent and accurate multi-gene signatures for classification of the four known subtypes of glioblastoma. Our results show that integrative application of feature selection and data discretization is crucial for successful platform transition and higher prediction accuracy of the derived molecular classifiers.

Keywords— isoform-level gene expression; data discretization; feature selection; platform transition; cancer subtype prediction

I. INTRODUCTION

High-throughput technologies have been extensively used for gene expression profiles. Application of these technologies has been accumulating invaluable data, particularly related to a disease. Indeed, the cancer genome atlas (TCGA) consortium offers unprecedented opportunity to derive robust and accurate gene-signatures for molecular classification [1]. The first prominent example is glioblastoma multiforme (GBM). Genomic profiling of GBM has allowed scientists to find biomarkers and categorize the disease into four subgroups. Despite the technological advances, analyzing and integrating the volumes of gene expression data from different platforms, however, remains a challenging task.

Computational approaches have been applied to derive multiclass classification models for patient stratification based on high-dimensional gene-expression [2]. Having fewer number of features in the final model is highly desirable not only from the machine learning perspective, but also in designing the validation experiments. For this purpose, feature selection methods are important in the analysis of high-dimensional data. In addition, discretization techniques can bridge the gap between different platforms for platform transition [3].

In conjunction with the preprocessing methods, several machine learning approaches have been applied to disease sample classification mostly focusing on one platform (e.g., microarray). Only recently, we developed PIGExClass [3] that captures and transfers gene signatures from one analytical platform to another. In this work, we explored four popular machine learning algorithms to predict the known subtypes of GBM patient samples.

II. METHODS

A. Dataset

Isoform-level gene expression estimates and molecular subtype information were obtained for 342 and 155 GBM samples profiled by Affymetrix exon arrays and RNA-seq, respectively. The four molecular subgroups are neural (N), proneural (PN), mesenchymal (M) and classical (CL). Gene expression profiles for 76 (18 are N; 22 are PN; 16 are M; and 20 are CL subtype) samples, available from both platforms, used to assess classification performance for platform transition. See [3] for more details.

B. Data format

We processed the gene expression data by fold change (FC), and two unsupervised discretization techniques, equal width (Equal-W) binning and equal frequency (Equal-F) binning, on the continuous FC data. Briefly, FC is a measure of a quantitative change of gene expression. Equal-W binning algorithm finds maximum and minimum values, and then divides the range into the user-defined equal width discrete intervals. Equal-F binning algorithm sorts all values in ascending order, and then divides the range into the user-defined intervals so that every interval contains the same number of sorted values.

C. Feature selection methods

SVM-RFE measures feature's importance to the data by iteratively eliminating one feature at a time [4]. The procedure is composed of training the SVM classifier, computing the ranking criteria w_i^2 for all features, and eliminating the feature with the lowest ranking criterion. This process is repeated until a small subset of features is achieved. RF based FS (RF based feature selection) uses both backward elimination strategy and the importance spectrum to search a set of important variables [5]. Concisely, multiple

978-1-4673-1321-6/12/$31.00 ©2014 IEEE

TABLE I. COMPARISON OF CLASSIFICATION METHODS TRAINED ON EXON-ARRAY (342 SAMPLES) AND TESTED ON RNA-SEQ (155 SAMPLES) TO PREDICT GBM SUBGROUPS. WE REPORT THE ACCURACY USING ALL AVAILABLE FEATURES AND THE BEST ACCURACY OF EACH CLASSIFIER IS MARKED IN BOLD.

Feature selection	**SVM-RFE**			**RF_based_FS**		
Classifier	**FC (%)**	**Equal-W (%)**	**Equal-F (%)**	**FC (%)**	**Equal-W (%)**	**Equal-F (%)**
SVM	52.7	50.0	**100.0**	51.4	36.9	98.7
RF	63.2	86.9	**97.4**	72.4	86.9	**97.4**
NB	34.3	35.6	85.6	39.5	34.3	**94.7**
PAM	47.4	38.2	88.2	36.9	29.0	**93.4**

random forests were iteratively constructed to search for a set of variable in each forest that yields the smallest out-of-bag (OOB) error rate.

D. Classification methods

SVM is a two-class classifier that constructs a hyperplane to separate the data with maximum margin [6, 7]. For the multiclass classification problem, we used a set of one-versus-one classifiers that determines the class by major voting. RF is an ensemble learning method based on decision trees with binary splits [8]. NB is a simple probabilistic classification method based on Bayes' theorem, used for calculating conditional probabilities, with an independence assumption [9]. PAM uses the nearest shrunken centroid method [10] that computes a standardized centroid for each class and then shrinks each of the class centroids by removing genes. A new sample is assigned to the nearest centroid.

E. Accuracy

We estimated the classification accuracy based on the number of correct predictions divided by the total number of prediction samples.

III. RESULTS

Using the classifiers that built on exon-array, we predicted the class labels of 155 TCGA samples based on RNA-seq data from which 76 samples are used to evaluate the classification accuracy.

Considering the features selected by SVM-RFE, SVM correctly predicted all the test examples followed by RF (97.4%), PAM (88.2%), and NB (85.5%) (Fig. 1A and TABLE I). Note that only SVM and RF achieved > 90% accuracy requiring about 600 genes or more. With no data discretization, SVM achieved only 52.7% accuracy.

With the features chosen from RF_based_FS, SVM slightly outperformed with an accuracy of 98.7% followed by RF, NB, and PAM with an accuracy of 97.4%, 94.7%, and 93.4%, respectively (Fig. 1B and TABLE I). However, without data discretization SVM achieved 51.4% accuracy while the best accuracy achieved by RF was 72.4%.

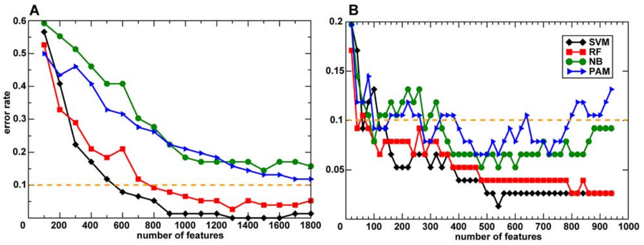

Fig. 1. Accuracy of classifiers with features from SVM-RFE **(A)** and RF_based_FS **(B)** with the Equal-F binning discretization. Exon-array (342 samples) and RNA-seq (155 samples) are used for training and testing, respectively. The orange dotted line marks the 90% accuracy.

IV. CONCLUSION

We presented an integrative application of feature selection and data discretization combined with the state-of-the-art machine leaning methods. Our analysis showed Equal-F binning led to higher accuracy of classification over FC and Equal-W binning when the model was trained on data from one platform and tested on the other platform. With Equal-F binning, RF_based_FS performed more efficiently than SVM-RFE. This is particularly obvious when fewer genes are considered in classification.

V. ACKKNOWLEDGEMENTS

The work was supported partially by the National Library of Medicine of the National Institutes of Health under Award Number R01LM011297. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

REFERENCES

[1] N. Cancer Genome Atlas Research, "Comprehensive genomic characterization defines human glioblastoma genes and core pathways," *Nature,* vol. 455, pp. 1061-8, Oct 23 2008.

[2] A. Statnikov, C. F. Aliferis, I. Tsamardinos, D. Hardin, and S. Levy, "A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis," *Bioinformatics,* vol. 21, pp. 631-43, Mar 1 2005.

[3] S. Pal, Y. Bi, L. Macyszyn, L. C. Showe, D. M. O'Rourke, and R. V. Davuluri, "Isoform-level gene signature improves prognostic stratification and accurately classifies glioblastoma subtypes," *Nucleic Acids Res,* Feb 6 2014.

[4] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research,* vol. 3, pp. 1157-1182, 2003.

[5] R. Diaz-Uriarte, "GeneSrF and varSelRF: a web-based tool and R package for gene selection and classification using random forest," *BMC Bioinformatics,* vol. 8, p. 328, 2007.

[6] B. Schölkopf, C. J. C. Burges, and A. J. Smola, "Advances in Kernel Methods," *The MIT Press,* 1998.

[7] V. Vapnik, "The Nature of Statistical Learning Theory," *Springer,* 1999.

[8] L. Breiman, "Random Forests," *Machine Learning,* vol. 45, pp. 5-32, 2001.

[9] T. M. Mitchell, "Machine Learning," *McGraw-Hill,* 1997.

[10] R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu, "Diagnosis of multiple cancer types by shrunken centroids of gene expression," *Proc Natl Acad Sci U S A,* vol. 99, pp. 6567-72, May 14 2002.