Classification of cancer subtypes using machine learning algorithms and biological data

Elsa Fecke[1], Mark Kon[2], Xinying Mu[3], Mia Li[4]

[1]St. Lawrence University; Boston University BRITE Bioinformatics, Summer 2016, [2,3,4]Department of Mathematics and Statistics, Boston University

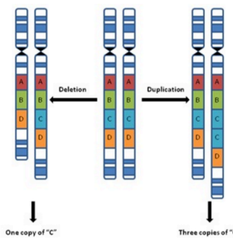
Machine Learning Algorithms

Machine learning (ML) methods for prediction of cancer subtype form a growing area of research in computational biology. Supervised ML algorithms make predictions of class membership (e.g., cancer subtype) based on provided evidence (e.g., gene expression vectors). On the other hand, unsupervised ML algorithms attempt to find patterns in the data from which predictions can be made without any prior information on class membership; clustering is an example of such an algorithm. ML methods can typically handle and take advantage of very high-dimensional data, such as gene expression (GE) and gene-level copy number (GLCN) data; these can be quite important in classification of cancer subtypes. The goal of this project is to determine how well GLCN data and GE data predict for cancer subtype, both separately and together. We benchmarked four supervised ML algorithms on six binary class GE datasets taken from Geman, *et al.*5 and used leave-one-out cross validation to benchmark the accuracy of these standard ML algorithms. The principal analysis was conducted on two datasets taken from Alvarez[1] that were partitioned for binary classification (see chart under results). Separately, the GE and GLCN data yield similar classification accuracies. The result of the combined dataset is still unknown. The combined dataset offers a variety of uses including not only cancer subtype prediction, but also predicted survival time and classification of cancer metastasis (Kim *et al.*7).

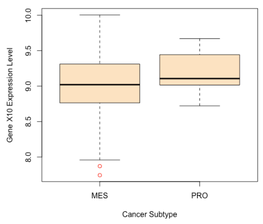
Introduction

Since cancer is a disease of the genome, using gene expression profiles has been a widely used approach to classification of cancer subtype. However, recent developments indicate that copy number variations (CNVs) are also important gene-level biomarkers of cancer subtype (Kim *et al.*[7]). CNV is an important biomarker in cancer given its extreme and consistent variations in different cancer subtypes. Normal cells have two copies of each gene; copy number variation occurs when a portion of a gene, or an entire gene, is duplicated or deleted in the process of carcinogenesis (Kim *et al.*[7]). Therefore, CNV has the potential to be as good as, and possibly better, at cancer subtype classification than gene expression. Synergizing these methods for cancer subtype classification may provide better diagnosis, prevention and treatment options. If subtype classification is optimized, then the appropriate drugs can be administered to patients with higher confidence.

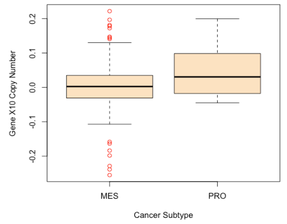
CNV[1]



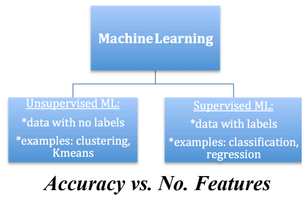
GE vs. Subtype



GLCN vs. Subtype

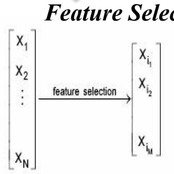


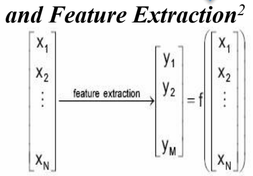
Machine Learning & Dimensionality Reduction



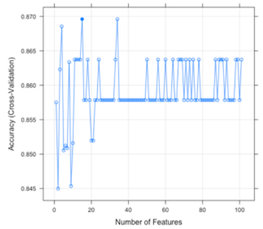
Reducing Dimensionality

Feature Selection and Feature Extraction[2]





Accuracy vs. No. Features



On the left: Feature selection optimizes cross-validation accuracy (CVA). The optimal CVA is often achieved by using a subset of features. Feature extraction also optimizes CVA, but instead of removing features, it transforms features to reduce dimensionality.

Machine Learning Algorithms

Naïve Bayesian Classifier (NBC)

The NBC classifies observations based on prior probabilities. The "naïve" assumption is that all features of the class variable (i.e. cancer subtype) are independent of each other. Let S be a sample with n gene expressions such that $S = (x_1, x_2, \ldots, x_n)$. Suppose there are two classes $a,b \in C$ where C is a class object.

The probability of S being some class C is: $$pr(C|S) = \frac{pr(S|C)pr(S)}{pr(S)}$$

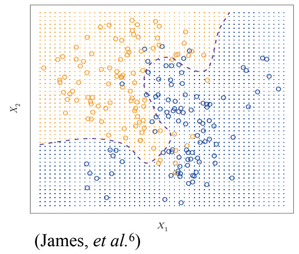
S is classified as a if and only if: $$\mathbb{C}_b(S) = \frac{pr(C=a|S)}{pr(C=b|S)} \geq 1$$

Since the Bayes classifier is "naïve", we assume: $$pr(S|C) = pr(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} pr(x_i|C)$$

Hence, the NBC is defined as: $$\mathbb{C}_{nb}(S) = \frac{pr(C=a)}{pr(C=b)} \prod_{i=1}^{n} \frac{pr(x_i|C=a)}{pr(x_i|C=b)}$$ (Zhang[9])

K-Nearest Neighbors Classifier (*k*-NNC)

Given a positive integer k, and a test observation X_0, the k-NN classifier identifies the number of k-neighboring points in the training set nearest to the observation X_0. In k-NN classification, the response is the class variable, so X_0 is assigned its class according to the majority vote of its k neighbors.

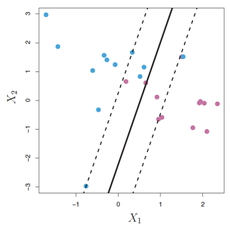


(James, *et al.*[6])

Support Vector Machine

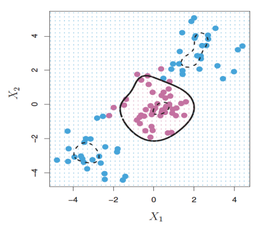
The Support Vector Machine (SVM) is an extension of the Support Vector Classifier (SVC). A SVC classifies observations based on a linear decision boundary, while a SVM can classify observations based non-linear decision boundaries (James, *et al.*[6]). SVM classifier was used in this project.

SVC



(James, *et al.*[6])

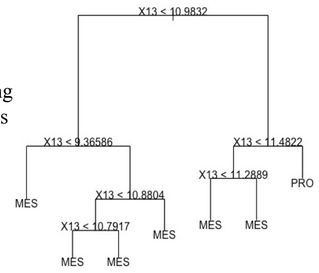
SVM with a radial basis function (RBF) kernel



(James, *et al.*[6])

Random Forest Classifier

To the right is a **classification tree** for gene X13.It predicts that each observation belongs to the most commonly occurring class in the training set in the node where it belongs. The class proportions in the training set of that particular node are also considered. A **random Forest** (RF) is a collection of trees in which each tree is a *different*, *random* subset of the training set: $\{A_i\}_{i=1}^{n}$ (Kon[8])



Footnotes

[1]A function that finds a way to compute the dot-product of points in high-dimensional space, without explicitly doing so (Ghose[10])

[2] One method of cross-validation that leaves out one sample, trains on n-1 samples and records a prediction. This process is executed n times to obtain n predictions which are then compared to the actual n results; this yields an LOOCV accuracy rate.

Acknowledgements

This work was funded by the Boston University Bioinformatics BRITE REU program

Results

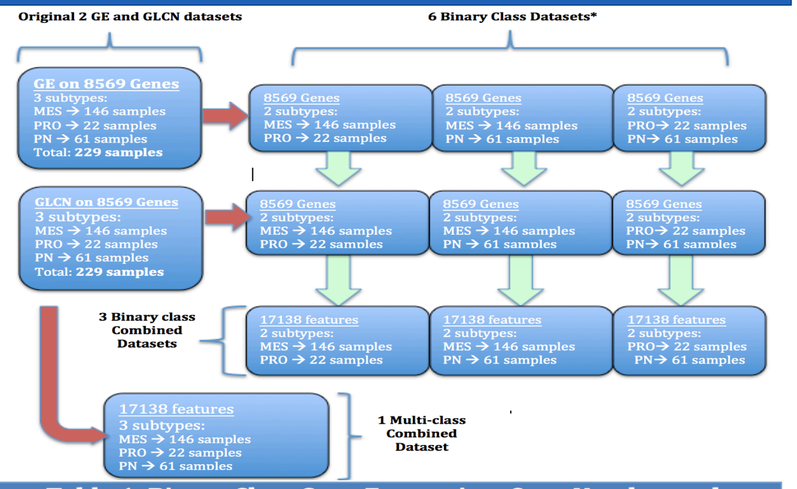


Table 1: Binary Class Gene Expression, Copy Number and Combined Datasets with Corresponding LOOCV Accuracies (%)

Dataset	No. features	No. MES samples	No. PRO samples	No. PN samples	NB	KNN	SVM	RF
MES & PRO	8569/ 8569 /17138	146	22	_	80.90/ 76.79/ 86.9	81.55/ 81.55/ 82.14	85.71/ 87.5/ 86.9	100/
MES & PN	8569/ 8569/ 17138	146	_	61	92.75/ 79.71/ 93.24	88.41/ 71.01/ 88.41	94.20/ 84.06/ 91.79	
PRO & PN	8569/ 8569/ 17138	_	22	61	79.52/ 65.06/ 78.31	75.9/ 65.06/ 78.31	77.11/ 71.08/ 75.9	

Conclusion

In this project, we have employed four ML algorithms including: KNN, NB, SVM and RF, on one GE cancer dataset and one GLCN cancer dataset. We benchmarked these algorithms on the two cancer datasets containing 3 subtypes and partitioned them into 3 pairs to perform binary classification. For the MES and PRO GE and GLCN dataset, the SVM classifier achieved the highest LOOCV accuracies at 85.7% and 87.5%, respectively. Additionally, within the MES-PRO datasets, the NB and SVM classifier performed similarly with a LOOCV accuracy of 86.9%. In general it appears that GE biomarker data is slightly better at predicting cancer subtype than GLCN biomarker data. The combined GE and GLCN data, in general, yield similar LOOCV accuracies to GE alone. There may be room for improvement, in terms of data structure, that could affect the outcome, but further analysis is required to confirm this. Possible next steps in supplementary analysis could include data manipulation as well as implementing new ML algorithms, such as clustering. This research has shown the combined GE and GLCN can produce impressive results as far as cancer subtype classification is concerned, but there are a number of alternative strategies that may optimize the LOOCV accuracies inthis type of combined data.

[1] Alvarez, M.J. (2014). Diggit: Inference of genetic variants driving cellular phenotypes. *R package version 1.4.0.*

[2] The approaches for reducing dimensionality [Digital image]. (n.d.). Retrieved from http://www.byclb.com/TR/Tutorials/neural_networks/ch5_1.htm

[3] Copy number variation [digital image]. (n.d). Retrieved from http://readingroom.mindspec.org/?page_id=8221

[5] Geman, D. *et al.* (2004) Classifying gene expression profiles from pairwise mRNA comparisons. *Statistical Applications in Genetic and Molecular Biology*, **3**, Article 19.

[6] James, G. *et al.* (2013) *An Introduction to statistical learning.* New York: Springer Science+Business Media.

[7] Kim, S. *et al.* (2015). A method for generating new datasets based on copy number for cancer analysis. *BioMed Research International,* **2015**, Article ID 467514

[8] Kon, M. (2016). Decision tree, random forests [PowerPoint]. Retrieved from Boston University Course named Mathematical and Statistical Methods of Bioinformatics

[9] Zhang, H. (2004) The optimality of naïve Bayes. *American Association for Artificial Intelligence.*

[10] Ghose, A. (2016, Feb, 24). Why does the RBF (radial basis function) kernel map into infinite dimensional space? [Web log post].Retrieved from https://www.quora.com/Why-does-the-RBF-radial-basis-function-kernel-map-into-infinite-dimensional-space