Cancer, SNPs and Machine Learning

Brett Poulin[1], Jennifer Listgarten[2], Russell Greiner[1], Sambasivarao Damaraju[2], Thomas Kolacz[1], Xiang Wan[1], David Wishart[3] and Brent Zanke[2*]

[1]Department of Computing Science, University of Alberta, Edmonton, [2]PolyomX, Cross Cancer Institute, Alberta Cancer Board, Edmonton and [3]Faculty of Pharmacy and Pharmaceutical Sciences, University of Alberta, Edmonton. *Corresponding Author: zanke@ cancerboard.ab.ca

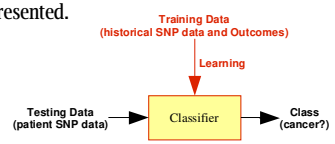
PolyomX.org

Background

SNPs(Single Nucleotide Polymorphisms) are commonly occurring genetic variations. SNPs may affect an individual's susceptibility to disease or response to particular drugs by altering the expression of the gene in which it occurs.

Cancer occurs through accumulation of mutations in multiple genes. The likelihood of mutagen-induced genetic alteration occuring and persisting may depend on efficiency of detoxification and repair capabilities. SNP variation may affect these processes.

Machine Learning deals with computer programs that learn from and improve with experience. (Mitchell, 1997) Learners are designed to recognize patterns in training data and classify new data as it is presented.

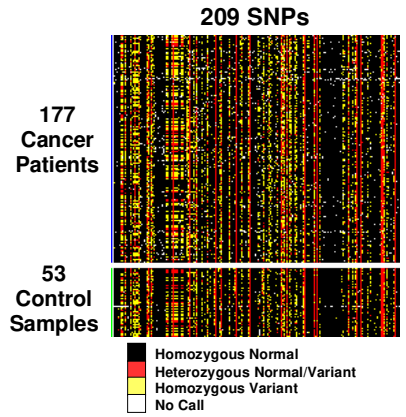


Studies which use machine learning across many SNPs with real clinical data are scarce.

Data

The experimental samples were obtained from 177 breast cancer patients from Edmonton and 53 control samples from an NIH Coriell panel.

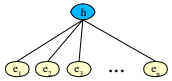
For each of the individuals sampled, 209 SNPs were chosen from across 68 genes. The genes chosen were drug metabolism genes, DNA repair enzymes, tumor suppressors, oncogenes, hormone receptors and signal transduction enzymes. Both synonymous and non-synonymous SNPs were chosen. Another control group is being acquired which is ethnically matched with the cancer group and will be compared with this preliminary study.



Methods

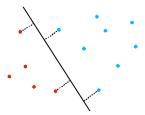
Patients were classified as having breast cancer or not using a number of machine learning techniques.

1. A naïve Bayes classifier seeks to find the most probable hypothesis given the evidence. It assumes independence of the SNPs.

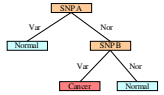


$$h = max_{h_i} P(h_i \mid e_i, e_2, \ldots, e_n)$$

2. A support vector machine (SVM) is used to find a linear separator between the cancer and non-cancer samples that minimizes the risk of errors. Although this may not be possible in a simple space representation, the data may be mapped to another feature space in which a separator may be found. (Vapnik, 1995)



3. Decision trees provide a decision flow diagram based on the training data. Test data is then classified based on the resulting tree.



Experiments were also conducted using artificial neural networks and clustering methods.

Results

All results were obtained using 5-fold cross-validation. The following matrices display the known classifications of the data against the predicted classifications for each learner.

1. naïve Bayes classifier

Known \ Predicted	Cancer	Normal	
Cancer	174	3	90% accuracy
Normal	20	33	

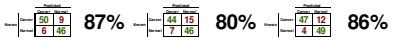
2. support vector machine

Known \ Predicted	Cancer	Normal	
Cancer	162	15	90% accuracy
Normal	9	44	

3. decision trees

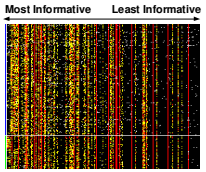
Known \ Predicted	Cancer	Normal	
Cancer	157	20	83% accuracy
Normal	18	35	

Since the baseline accuracy (ZeroR) of the unbalanced data set is 77%, we also made subsamples of the cancer group and reran each against the controls. The results below are for naïve Bayes with baseline ~53%.



The SNPs may also be ranked by measures of importance, including information content and p-value from a Chi-squared test.

209 SNPs sorted by Information Content



Sorting by p-value results in similar ordering.

Conclusions

Differences in SNP profiles between sample groups can be recognized through the use of machine learning techniques. These statistical techniques also give a framework in which the relative contribution of each SNP to the outcome can be assessed.

The biological significance of these SNP variations with respect to cancer prediction remains to be resolved pending better understanding of the impact of control design in SNP studies(Wacholder *et al.*). Further analysis with a larger group of ethnically matched controls will address this issue in the near future. This preliminary analysis demonstrates the utility of machine learning techniques in discriminating between populations based on real SNP data.

Acknowledgements

Polyomx (www.polyomx.org) is a cancer research initiative based at the Cross Cancer Institute, Edmonton, Alberta, Canada.

Brett Poulin (poulin@cs.ualberta.ca) was partially funded by PENCE.

References

Mitchell, T. *Machine Learning.* McGraw-Hill, Boston, 1997.

Vapnik, V. *The Nature of Statistical Learning.* Springer, New York, 1995.

Witten, I. And Frank, E. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations.* Morgan Kaufmann, 1999. (WEKA)

Joachims, T. Making large-scale SVM learning practical. *Advances in Kernel Methods - Support Vector Learning.* Scholkopf, B., Burges, C. and A. Smola (ed.), MIT-Press, 1999. (SVM^{light} Software)

Wacholder, S., Rothman, N. and Caporaso, N. Population stratification in epidemiologic studies of common genetic variants and cancer: Quantification of Bias. *Journal of the National Cancer Institute.* 92:1151-1158. 2000.