Predicting Breast Cancer and Prostate Cancer Susceptibility from Single Nucleotide Polymorphisms

Jie Liu JIELIU@CS.WISC.EDU
Department of Computer Sciences, UW-Madison

Elizabeth Burnside EBURNSIDE@UWHEALTH.ORG
Department of Radiology
Department of Biostatistics and Medical Informatics, UW-Madison

David Page PAGE@BIOSTAT.WISC.EDU
Department of Biostatistics and Medical Informatics
Department of Computer Sciences, UW-Madison

Abstract

Large-scale genome-wide genetic profiling using markers of SNPs provides opportunities to investigate the possibility of using those biomarkers for predicting genetic risks. Recent computational studies have identified some associated genetic variations which can explain a fraction of breast cancer risk and prostate cancer risk. We attempt to build accurate classification models for predicting disease susceptibility based on human SNPs. We firstly carry out feature selection via logistic regression coupled with a likelihood ratio test and remove a large number of irrelevant SNPs. Then, we employ supervised learning method SVM to build classification models. Our computational results show that our feature selection method based on logistic regression and likelihood ratio test can effectively select relevant features for SVM on the prostate cancer dataset, whereas it does not help SVM very much when applied on the breast cancer dataset.

1. Introduction

Familial aggregation of some diseases such as breast cancer and prostate cancer demonstrates there exist genetic factors explaining the risk to some extent. Large-scale genome-wide genetic profiling using markers of SNPs provides good opportunities to investigate the possibility of using those biomarkers for predicting disease risks, and results in the great boom of genome wide association studies on different heritable diseases (Hunter et al., 2007; Yeager et al., 2007). In statistics and machine learning terminology, typically an example in GWAS is a human, the class variable is a disease such as breast cancer or prostate cancer, and the features are single positions in the entire genome where individuals can vary, known as single nucleotide polymorphisms (SNPs). One goal in GWAS is to find a subset of SNPs that can be used to predict the class variable. In addition, the locations of the predictive SNPs on the genome can give insight into the genetics of the disease. The identification of the interactions between those SNPs can also lead to a better understanding of the disease.

The human genome has roughly three billion positions, roughly three million of which are SNPs. At the time of this writing, state-of-the-art technology enables measurement of a million SNPs in one experiment for a cost around 300 US dollars. Although this means the full set of known SNPs cannot be measured, SNPs that are close together on the genome are often highly correlated. Hence the omission of some SNPs is not as much of a problem as one might first think. Instead, we have the problem of strong-correlation among our features: most SNPs are very highly correlated with one or more nearby SNPs, with R^2 (squared correlation coefficient) values well above 0.8.

The primary problem of GWAS is that there are much more predictor variables than samples, namely the

Appearing in ICML 2013 Workshop on Role of Machine Learning in Transforming Healthcare.

$p \gg n$ problem in statistics or the curse of dimensionality in machine learning. In those cases, feature and variable selection (Guyon & Elisseeff, 2003) is one necessary step before we can build the supervised learning models. However, variable selection methods such as forward selection or backward deletion will not work because they are usually computation intensive. Therefore, we need to carry out variable selection more efficiently. We build a Logistics regression model for each SNP, and test the fitness of the model with likelihood ratio test, and use the P-values from the tests to rank the SNPs.

In addition, another challenge of GWAS is that there exists strong correlation between predictor variables. Therefore, we do not use generalized linear models. Instead, we employ support vector machine to build the classification model because support vector machine can be expected to work better in this circumstance.

Our computational results showed that the feature selection method based on logistic regression and likelihood ratio test can effectively select relevant features for the classification algorithm SVM on the prostate cancer dataset. However, the feature selection method does not help SVM very much when applied on the breast cancer dataset. One explanation is that genetic factors can explain different amount of risk for different diseases. It is estimated that 27% breast cancer is caused by genetics whereas about 42% prostate cancer is caused by genetics (Lichtenstein et al., 2000). Therefore, we can expect the feature selection and supervised learning for breast cancer is more difficult than prostate cancer.

2. Datasets

Our datasets in the experiment are from National Cancer Institutes Cancer Genetics Markers of Susceptibility project [1]. The breast cancer dataset contains 528,173 SNPs (genotyped by Illumina HumanHap500) from 1145 breast cancer patients and 1142 controls with European ancestry. The prostate cancer dataset contains 546,593 SNPs from 1176 prostate cancer patients and 1101 controls (see Table 1).

Table 1. Description of the datasets.

	♯ SNPs	♯ Cases	♯ Controls
Breast Cancer	528,173	1,145	1,142
Prostate Cancer	546,593	1,176	1,101

The original dataset records the nucleic acids appearing at the locus on chromosomes of every single sample. In our experiment, we convert AA into 1, AB into 0, and BB into -1 where A stands for the common allele at this locus and B stands for the minor allele. This encoding method has been used in many studies such as the work of Waddell et al. (2005).

[1] https://caintegrator.nci.nih.gov/cgems/

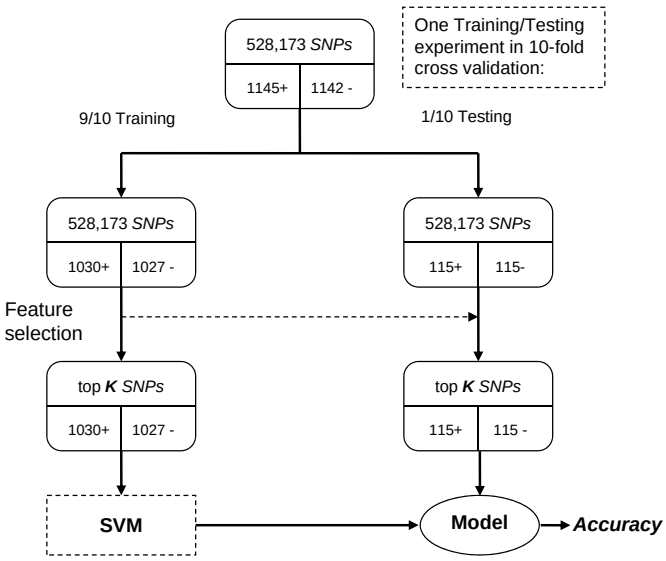


Figure 1. Flow chart of the experiments.

3. Methods

Due to the extremely high dimensionality, we cannot use all the features to train our model. We firstly carry out variable selection on the training dataset to remove non-relevant features. With a much smaller feature set, we employ SVM to build the classification models and evaluate the models on the testing dataset. To evaluate them in a fair way, we repeat the experiments in 10-fold cross validation fashion. The whole procedure of feature selection and building classification models is showed in Figure 1.

So far, a wide variety of feature selection approaches has been proposed and can be classified into two categories, namely filters and wrappers (Guyon & Elisseeff, 2003). Usually filters assume independence between features and rank them individually according to some relevance criterion. Wrapper methods iteratively generate a candidate feature subset and test it by a specific learning algorithms performance, until some criterion is satisfied. Wrapper methods are usually much more computation intensive. Therefore we use one typical filter method, namely logistic regression coupled with likelihood ratio test.

Logistic regression is one generalized linear model for

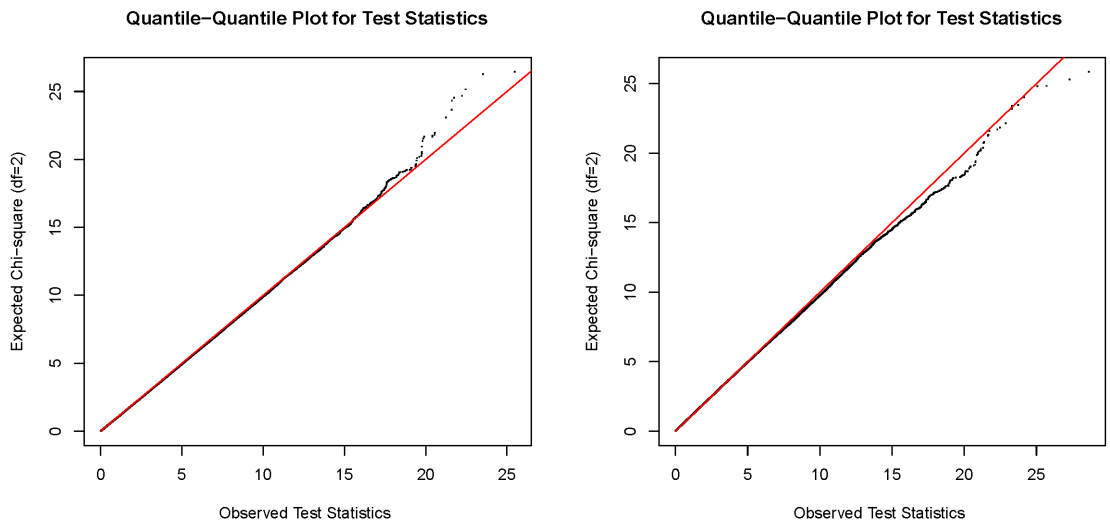


Figure 2. QQ plots for the likelihood ratio test statistics in the breast cancer data (left) and prostate cancer data (right).

binary response variables, which models the log odds ratio as a linear model of predictor variables. In our problem, we build a logistic regression model for each SNP, in which we code one SNP as two indicator variables for heterozygous carrier of the minor allele (X_1) and homozygote carrier of the minor allele (X_2). In other words, we convert AA into $X_1 = 0$, $X_2 = 0$, AB into $X_1 = 1$, $X_2 = 0$, and BB into $X_1 = 0$, $X_2 = 1$ where A stands for the common allele at this locus and B stands for the minor allele.

One way of testing the models fitness is the likelihood ratio test. Under the null hypothesis, the test statistic S has an asymptotic chi-square distribution with q degrees of freedom. In our problem, q is 2. The p-values from the tests not only tell us the fitness of the logistic regression model, but also tell us the relevance of the SNPs with the disease. The smaller the p-value is, the corresponding SNP is more highly correlated with the disease. With this criterion, we pick the top 50, top 100, top 150, top 200, ..., top 1,000 features and use them in different independent experiments so as to evaluate how the feature size affects these computational supervised learning methods.

In our experiment, we use one popular supervised learning algorithm SVM to generate our classification models. Like other classification tasks, we evaluate our classification models based on accuracy.

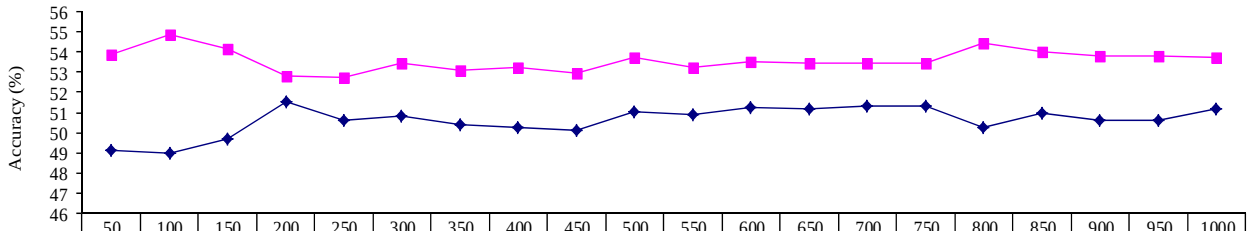
4. Results and Discussion

In order to demonstrate the existence of associated SNPs, we make the quantile-quantile plots for the test statistics in the likelihood ratio test with two degrees of freedom (shown in Figure 2). Figure 2(a) is for breast cancer and Figure 2(b) is for prostate cancer. If all the SNPs are not associated with the diseases, we expect the plots line up perfectly on the 45 degree straight line (in red). However, in both of the two diseases, we can observe a clear deviation from the straight line if the statistics is larger enough. Another interesting observation is that for the two diseases, the quantile-quantile plots depart from the straight line to different directions.

Figure 3 shows the prediction performance of SVM with different numbers of SNPs selected on the breast cancer dataset and the prostate cancer dataset. The accuracies are average accuracies from 10-fold cross validation. For breast cancer, the feature selection method does not help very much to improve the classification performance for SVM. When we are using the top 50 SNPs, the accuracy is 49.1%. As we add more SNPs, the accuracy improves a little. When we use 200 SNPs, SVM's classification accuracy peaks at 51.55%. After that, adding more SNPs does not help improving the performance.

When our method is applied to the prostate cancer dataset, the overall performance is better than the performance on breast cancer. When the top 100 SNPs are selected, SVM achieves its best performance with classification accuracy 54.88%. When we use more SNPs, the performance does not get improved and the corresponding accuracies are between 53% and 54%.

One explanation of the performance difference between the two diseases is that most diseases have both a ge-



	50	100	150	200	250	300	350	400	450	500	550	600	650	700	750	800	850	900	950	1000
Breast Cancer	49.1	48.97	49.72	51.55	50.63	50.81	50.38	50.24	50.11	51.03	50.89	51.25	51.16	51.29	51.29	50.24	50.98	50.59	50.64	51.16
Prostate Cancer	53.87	54.88	54.17	52.81	52.77	53.48	53.08	53.26	52.99	53.74	53.26	53.56	53.43	53.48	53.43	54.44	54.05	53.78	53.82	53.74

Figure 3. Performance of SVM against the number of SNPs selected on breast cancer data and prostate cancer data.

netic and environmental component, and the genetic factors can explain different amount of disease risk. It is estimated that 27% breast cancer is caused by genetics whereas about 42% prostate cancer is caused by genetics (Lichtenstein et al., 2000). Therefore, if we just use the genetic factors to build the classification model, we could expect the model for prostate cancer works better than the model for breast cancer.

Future work is additional empirical study. First, it would be useful to repeat the comparative experiments in the present paper with other GWAS data sets. Second, it also would be interesting to propose and implement new feature selection algorithms which could potentially remove the redundant features. Finally, we would like to explore using wrapper-based feature selection algorithms. While wrapper-based approaches are computationally infeasible on the original high-dimension data, once our feature selection method has been used to filter out a big fraction of irrelevant features, a wrapper-based approach could be employed for further feature subset selection.

Acknowledgements

The authors gratefully acknowledge the support of NCI grant R01CA127379-01 and its ARRA supplement 3R01CA127379-03S1, NLM grant R01LM010921, NCI grant R01CA165229, NIGMS grant R01GM097618-01 and NLM grant R01LM011028-01.

References

Guyon, Isabelle and Elisseeff, André. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

Hunter, David J., Kraft, Peter, Jacobs, Kevin B., Cox, David G., Yeager, Meredith, Hankinson, Susan E., Wacholder, Sholom, Wang, Zhaoming, Welch, Robert, Hutchinson, Amy, Wang, Junwen, Yu, Kai, Chatterjee, Nilanjan, Orr, Nick, Willett, Walter C., Colditz, Graham A., Ziegler, Regina G., Berg, Christine D., Buys, Saundra S., Mccarty, Catherine A., Feigelson, Heather S., Calle, Eugenia E., Thun, Michael J., Hayes, Richard B., Tucker, Margaret, Gerhard, Daniela S., Fraumeni, Joseph F., Hoover, Robert N., Thomas, Gilles, and Chanock, Stephen J. A genome-wide association study identifies alleles in fgfr2 associated with risk of sporadic postmenopausal breast cancer. *Nature Genetics*, 39(7):870–874, 2007.

Lichtenstein, Paul, Holm, Niels V., Verkasalo, Pia K., Iliadou, Anastasia, Kaprio, Jaakko, Koskenvuo, Markku, Pukkala, Eero, Skytthe, Axel, and Hemminki, Kari. Environmental and heritable factors in the causation of cancer–analyses of cohorts of twins from sweden, denmark, and finland. *New England Journal of Medicine*, 343:78–85, 2000.

Waddell, Michael, Page, David, Zhan, Fenghuang, Barlogie, Bart, and Shaughnessy, Jr., John. Predicting cancer susceptibility from single-nucleotide polymorphism data: A case study in multiple myeloma. In *Proceedings of BIOKDD '05*, 2005.

Yeager, Meredith, Orr, Nick, Hayes, Richard B, Jacobs, Kevin B, Kraft, Peter, Wacholder, Sholom, Minichiello, Mark J, Fearnhead, Paul, Yu, Kai, Chatterjee, Nilanjan, et al. Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nature genetics*, 39(5):645–649, 2007.