Bag of Naïve Bayes: biomarker selection and classification from Genome-Wide SNP data

Francesco Sambo, Emanuele Trifoglio, Barbara Di Camillo, Gianna Toffolo and Claudio Cobelli
Department of Information Engineering, University of Padova, Italy

Introduction

In the past few years, the hereditary component of complex multifactorial diseases has started to be explored through the novel paradigm of Genome Wide Association Studies (GWASs). A GWAS searches for patterns of genetic variation, in the form of Single Nucleotide Polymorphisms (SNPs), between a population of affected individuals (cases) and a healthy control population. The objective of a GWAS is twofold: on the one hand, one searches for the set of SNPs that best explains the hereditary component of the disease (genetic biomarkers), on the other hand, one tries to learn a rule for classifying unknown subjects as cases or controls, given their genetic profile and possibly other environmental covariates.

Multifactorial diseases have an heterogeneous nature, arising from complex patterns of interaction between a set of genetic traits and the environment: to fully capture the optimal set of genetic biomarkers, thus, all SNPs in a GWAS should be analyzed simultaneously [4]. The extremely large numbers involved (O(10^6) SNPs in O(10^3) individuals), however, have led the vast majority of studies to rely upon single SNP association tests [5].

Genetic linkage, i.e. the non random association between portions of the genome close to each other, acts as a confounding factor for the analysis: in the proximity of a true causal biomarker, several SNPs both highly correlated with the biomarker and mildly associated to the disease are often observed in GWASs.

In this work, we present Bag of Naïve Bayes (BoNB), an algorithm for classification and biomarker selection from Genome Wide SNP data. Our algorithm is based on Naïve Bayes (NB) classification, powered by three additional strategies: (a) a bagging of Naïve Bayes classifiers, to improve robustness of the predictions, (b) a novel strategy for ranking and selecting the attributes used by each bagged classifier, to enforce attribute independence, and (c) a permutation-based procedure for selecting significant biomarkers, based on their marginal utility in the classification process.

BoNB was tested on the WTCCC case-control study on Type 1 Diabetes [5]. Classification accuracy, assessed through 10-fold cross validation and measured with the Matthews Correlation Coefficient [1], proved significantly higher than the ones obtained by both a standard Naïve Bayes classifier, trained on the SNPs that reached genome-wide significance in a single SNP test, and by HyperLasso, a state-of-the-art penalized logistic regression technique specifically designed for the simultaneous analysis of genome-wide data [4].

Methods

Given a dataset X={X_1…X_n}, consisting of p observations (subjects) of n attributes (SNPs), and a set Y of class labels for each observation (case/control), a Naïve Bayes Classifier (NBC) estimates from the dataset a classification rule in the form

\[
Pr(Y = y_k | X_1…X_n) = \frac{Pr(Y = y_k) \prod_{i} Pr(X_i | Y = y_k)}{\sum_{j} Pr(Y = y_j) \prod_{i} Pr(X_i | Y = y_j)}.
\]

For discrete-valued variables, such as SNPs, probability distributions Pr(Y = y_k) and Pr(X_i | Y = y_k) are represented with conditional probability tables, which are estimated from the data by counting the occurrences of each combination of attribute values and class labels.

Our algorithm, Bag of Naïve Bayes (BoNB), consists in an ensemble of Naïve Bayes Classifiers, trained on GWAS data with the procedure known as Bootstrap Aggregating or Bagging [2].
The Bagging procedure starts by computing a set of Bootstrap samples of the dataset $X$, i.e. a set \( \{X^{(1)} \ldots X^{(b)}\} \) of replicated datasets, each one obtained by sampling $p$ observations with replacement from the original set $X$. A Naïve Bayes Classifier $NBC^{(b)}$ is then trained on each Bootstrap sample $X^{(b)}$. Classification probability is then obtained by averaging the output class probabilities estimated by each $NBC^{(b)}$.

For estimating probabilities as in Eq. (1), the Naïve Bayes Classifier makes the assumption that the attributes $X_1 \ldots X_r$ are all conditionally independent of one another, given $Y$. Such an assumption is unlikely to hold if all the SNPs of a GWAS are exploited as attributes for the NBCs, because of genetic linkage. Moreover, computing Eq. (1) for the whole SNP set can be computationally cumbersome and can lead to numerical and overfitting problems.

We thus developed a procedure for ranking and selecting a good set of independent attributes for each $NBC^{(b)}$. For each Bootstrap sample $X^{(b)}$, the procedure first ranks each SNP according to the $p$-value of the test is lower than 0.05. MCC ranges from -1 (all examples incorrectly classified) to 1 (all correctly classified) and equals 0 in case of majority classification, i.e. when all labels are assigned to the most represented class.

From the ranked list, SNPs are iteratively added as attributes of $NBC^{(b)}$ in decreasing order of estimated MCC. Every time a SNP is included as an attribute, all SNPs in the ranked lists which are both close to the included SNP on the genome (distance < 1Mb) and correlated with it ($r^2 > 0.1$) are deleted from the list: this enforces attribute independence, thus coping with the problems arising from genetic linkage. SNPs are added in groups of exponentially increasing size, as long as the generalization ability of $NBC^{(b)}$ increases: to estimate the generalization ability, we test $NBC^{(b)}$ on the Out-of-Bag sample $OOB^{(b)}$, consisting of all the observations left out from $X$ when sampling $X^{(b)}$, and measure the MCC of the prediction. The exponential increase in the number of added attributes allows BoNB to reach the adequate size for the attribute set of each NBC in a logarithmic number of steps.

Such a procedure, iterated for the $B$ bootstrap samples, results in an ensemble of $B$ Naïve Bayes Classifiers, each with a possibly different set of features. New, unseen subjects are then classified by having each NBC estimate output class probabilities and by averaging the probabilities across all $B$ NBCs. Such an approach is known for increasing the robustness of the predictions [2].

For the second objective of GWASs, biomarker selection, we adapted for BoNB a procedure originally designed for the Random Forests bagged classifier [3]: for each of the SNPs included as attributes by at least one NBC, we randomly permute the subject labels, test each $NBC^{(b)}$ on its corresponding $OOB^{(b)}$ and record the relative decrease in MCC due to the permutation. Such a measure can be used as an indicator of the relative importance of each attribute, given all the other attributes of the NBC. For each SNP, the permutation procedure returns a list of values of relative decrease in MCC, one value for each NBC that included the SNP: we test the significance of the decrease with a one-tailed Wilcoxon test on the list of values, marking as biomarkers the SNPs for which the $p$-value of the test is lower than 0.05.

BoNB was tested on the WTCCC case-control study on Type 1 Diabetes [5], consisting of 458376 SNPs measured for 1963 T1D cases and 2938 healthy controls (after the application of all Quality Control filters reported in [5]). Predictive accuracy of BoNB with $B=100$ NBCs was estimated with 10-fold cross validation and measured with MCC. Accuracy was compared with the ones obtained by a simple Naïve Bayes Classifier, trained on all the SNPs that reached the significance threshold
of $5 \times 10^7$ ([5]) in a single 2df $\chi^2$ test of association with a general genetic model, and by HyperLasso, a state-of-the-art logistic regression method for the simultaneous analysis of all SNPs in a Genome-Wide association study.

**Results**

On the experimental dataset, BoNB reached an MCC of 0.55±0.02 (mean ± standard deviation), significantly higher than the ones reached by both the simple Naïve Bayes Classifier (0.31±0.05, Wilcoxon signed-rank p-value 0.002) and by HyperLasso (0.50±0.03, p-value 0.0089).

For biomarker selection, we run BoNB on the whole dataset and compared its results with the biomarkers identified by HyperLasso and by the general 2df test. The average number of attributes included by BoNB in each NBC was 2.94 and 35 SNPs were included by at least one NBC, mainly from the MHC region of chromosome 6, but also from chromosomes 1 (gene RSBN1), 3 (EPHA3 and CCR2), 7 (CHN2), 8 (SOX17), 12 (NAA25) and 16 (IRX3). Among these, only SNPs rs9273363, rs3101942, rs492899, rs6936863, rs805301, rs9275418 and rs2856688 on chromosome 6 (MHC region) and SNP rs6679677 on chromosome 1 (gene RSBN1) reached the significance level on the permutation test.

Compared to the 394 SNPs that reached the significance level on the 2df general test and to the 180 SNPs returned by HyperLasso, the list of biomarker selected by BoNB is more compact, but it does not prevent BoNB to reach significantly higher classification accuracy.

**Discussion**

Learning an ensemble of classifiers from a bootstrap sample of the original datasets provides BoNB with two main advantages: on the one hand, it guarantees a higher generalization ability in classification by increasing the stability of the learning process [2], on the other hand, it allows to select biomarkers in a sound and statistically principled way.

Two features of the Naïve Bayes Classifier make it rather appealing for Genome-Wide data analysis: on the one hand, conditional probability table analysis does not assume a pre-specified model of genetic effect, on the other hand, missing values are seamlessly handled by both the learning and the classification procedure.

We conclude with a note on computational performance: BoNB is implemented in C++ and, on the WTCCC dataset, takes approximately 20 minutes for training 100 NBCs and selecting the biomarkers. A careful allocation strategy makes BoNB occupy around 600 MB of RAM, allowing it to be easily run on a desktop computer.

**Acknowledgements**

The research was supported by the European Union's Seventh Framework Programme (FP7/2007-2013) for the Innovative Medicine Initiative under grant agreement n° IMI/115006 (the SUMMIT consortium). This study makes use of data generated by the Wellcome Trust Case-Control Consortium. A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk. Funding for the project was provided by the Wellcome Trust under award 076113 and 085475.

**References**

[1] Baldi P, Brunak S, Chauvin Y, Andersen CAF, Nielsen H. Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics 16, 412–424, 2000.

[2] Breiman L. Bagging predictors. Machine Learning 24 (2): 123–140, 1996.

[3] Breiman L. Random Forests. Machine Learning 45 (1): 5–32, 2001.

[4] Hoggart CJ, Whittaker JC, De Iorio M, Balding DJ. Simultaneous Analysis of All SNPs in Genome-Wide and Re-Sequencing Association Studies. PLoS Genet. 4(7): e1000130, 2008.

[5] Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 447:661-78, 2007.