A Hierarchical Naïve Bayes Strategy for Genetic Association Studies

MALOVINI A (1, 2), BARBARINI N (1), BELLAZZI R (1)
(1) Department of Computer Engineering and Systems Science - University of Pavia, Italy
(2) Salvatore Maugeri Foundation, Pavia, Italy

Motivation
Genome Wide Association Studies (GWAS) represent powerful approaches that aim at disentangling the genetic and molecular mechanisms underlying complex traits. The usual ‘one-SNP-at-the-time’ testing strategy cannot capture the multi-factorial nature of this kind of disorders. We propose here an extension of the classical Naïve Bayes (NB) classification model, called Hierarchical Naïve Bayes (HNB), for taking into account associations regarding SNPs that map to a certain chromosome region (i.e. genic locus or haplotype block). Five folds cross validation showed that our model reaches classification performances that were higher with respect to the standard Naïve Bayes classifier (NB) for both simulated and real datasets. Further, the generalization capability obtained by the HNB algorithm using a haplotype-based SNPs grouping schema was higher than those obtained by using standard classification approaches based on individual haplotype-phases as predictors.

Methods
In the HNB implemented, the SNPs mapping to the same gene are considered as ‘details’ or ‘replicates’ of the locus; each contributes to the overall effect of the gene on the phenotype. A latent variable for each region, which models the ‘population’ of correlated SNPs, can be then used to summarize the available information. The classification is thus performed relying on the HBN conditional probability distributions and on the SNPs data available. The HNB algorithm implemented extends the classifier previously described by Demichelis et al [1] to discrete distributions, relying on a multinomial-Dirichlet model to describe the data and the (population) latent variable, while the learning step is performed with the ‘collapsing’ strategy presented by Bellazzi et al [2].

Results
The developed methodology has been validated on 3 simulated datasets, each composed by 300 cases, 300 controls and a variable number of genic SNPs, designed to reflect the patterns of allele frequency and linkage disequilibrium of the general HapMap caucasian population. Further, our approach has been applied to a real genome-wide dataset composed by 400 cases, 400 controls and 177 SNPs mapping to 9 genes derived from a GWAS on Type 2 Diabetes (T2D) by the Wellcome Trust Case
Control Consortium (WTCCC). Five folds cross validation using a gene-based SNPs grouping strategy showed that our approach reaches better classification accuracies on both simulated and real data with respect to the standard NB (sim 1: 0.68 [0.61-0.75] vs 0.62 [0.54-0.69]; sim 2: 0.77 [0.70-0.83] vs 0.74 [0.67-0.80]; sim 3: 0.81 [0.74-0.86] vs 0.73 [0.66-0.79] and WTCCC: 0.66 [0.60-0.71] vs 0.65 [0.60-0.71]). Also, the results obtained by haplotype-based SNPs grouping strategy on the same datasets suggest that the proposed HNB approach is able to reach better or equal classification accuracies than the usual haplotype-phases based approach (sim 1: 0.81 [0.77-0.85] vs 0.78 [0.74-0.8]; sim 2: 0.87 [0.83-0.91] vs 0.83 [0.80-0.87]; sim 3: 0.88 [0.85-0.92] vs 0.84 [0.80-0.88] and 0.75 [0.60-0.90] vs 0.75 [0.60-0.90]). The approach proposed in this abstract allows to deal with classification of examples for which SNPs measurements are available for each chromosome region used as reference for grouping related SNPs. HNB improves the NB performances by properly handling the within-loci variability in terms of predictive capability. Finally, one of the main advantages of this approach is represented by its capability to resume the information from a set of structurally (haplotypes) and functionally related (genes) markers.

**Acknowledgement**

Funding for the project was provided by the Wellcome Trust under award 076113.

**Supplementary information**

This study makes use of data generated by the WTCCC. A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk.

**References**


Contact email
alberto.malovini@unipv.it