DEPARTMENT OF
INFORMATION
ENGINEERING
UNIVERSITY OF PADOVA

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

Francesco Sambo

Context

Complex disease, with hypothesized but still not understood genetic origin

⬇

Genome Wide Association Study (GWAS)

- O(10^6) Single Nucleotyde Polymorphisms (SNPs)
- O(10^3) case / control individuals

Objectives:

1. Biomarker Selection
2. Classification

Bag of Naïve Bayes (BoNB)

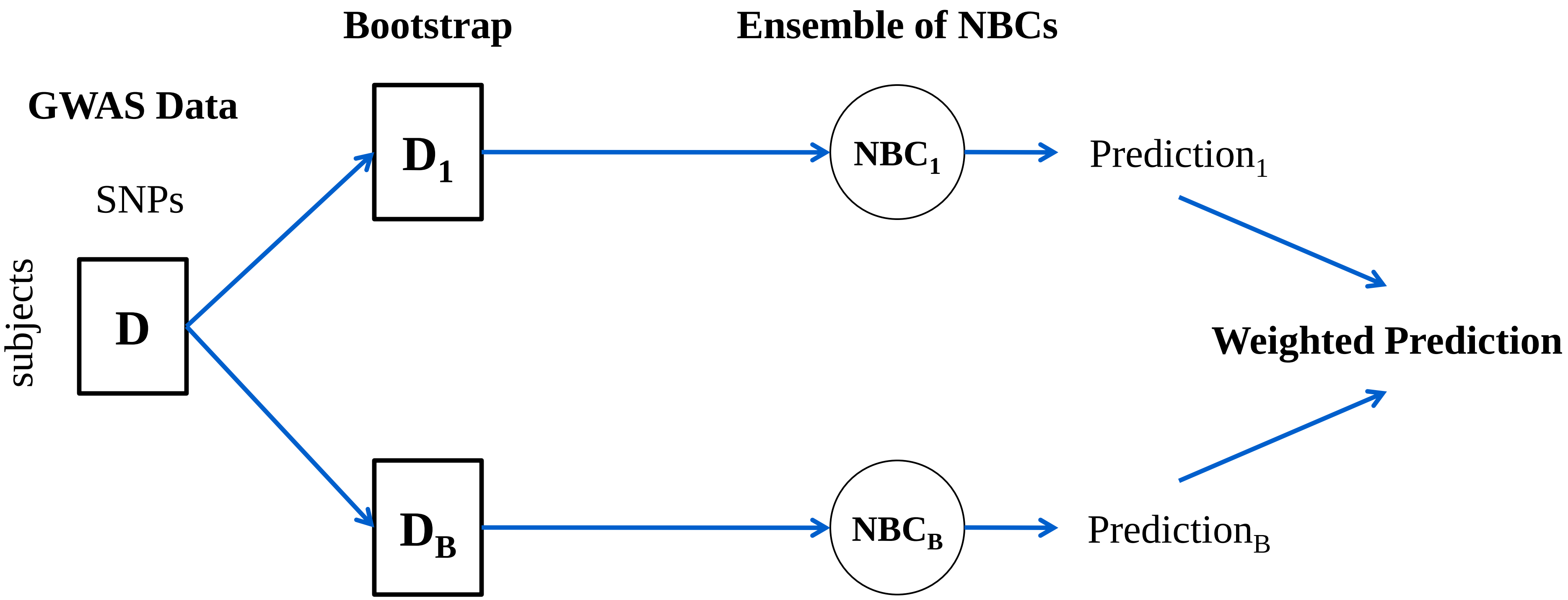
- Both classification and biomarker selection
- Based on Naïve Bayes classification
- Main features:

a) Ensamble of Naïve Bayes Classifiers (NBC), robustness

b) Novel strategy for ranking and selecting attributes for each NBC, attribute independence

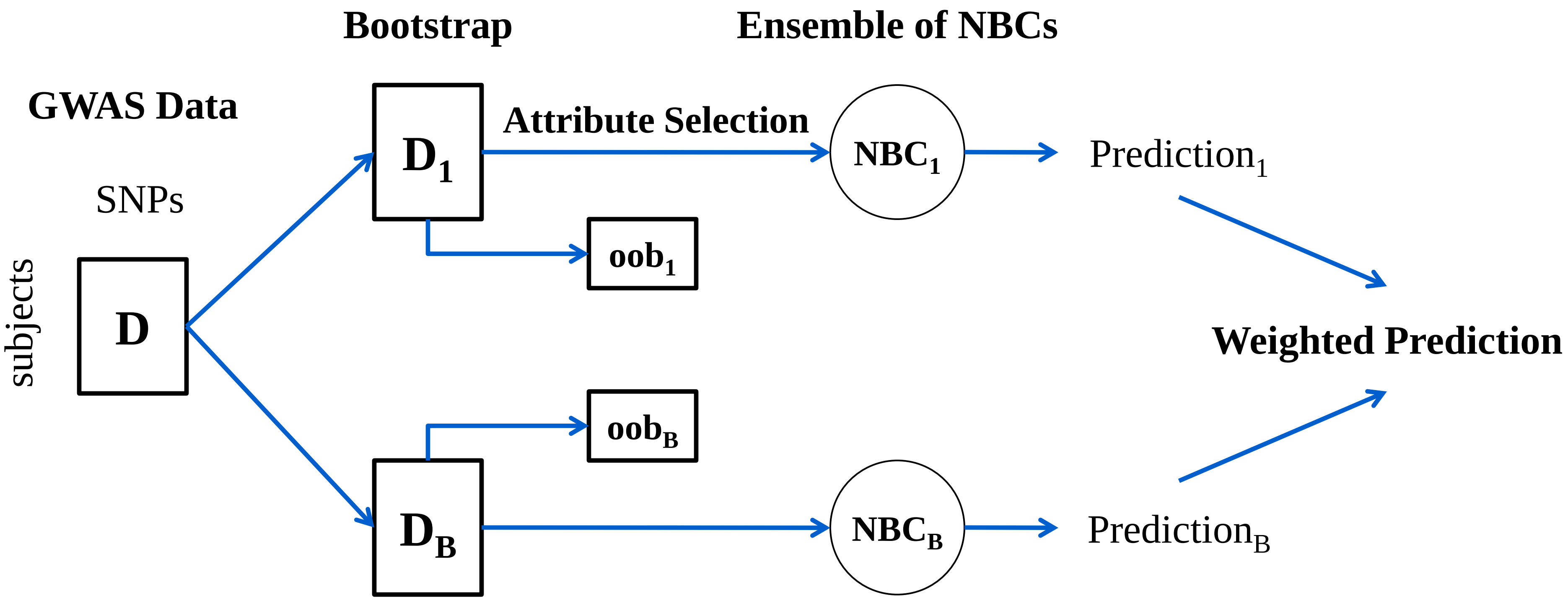
c) Permutation-based procedure for biomarker selection, based on marginal utility.

Bagging (Bootstrap AGGregatING)



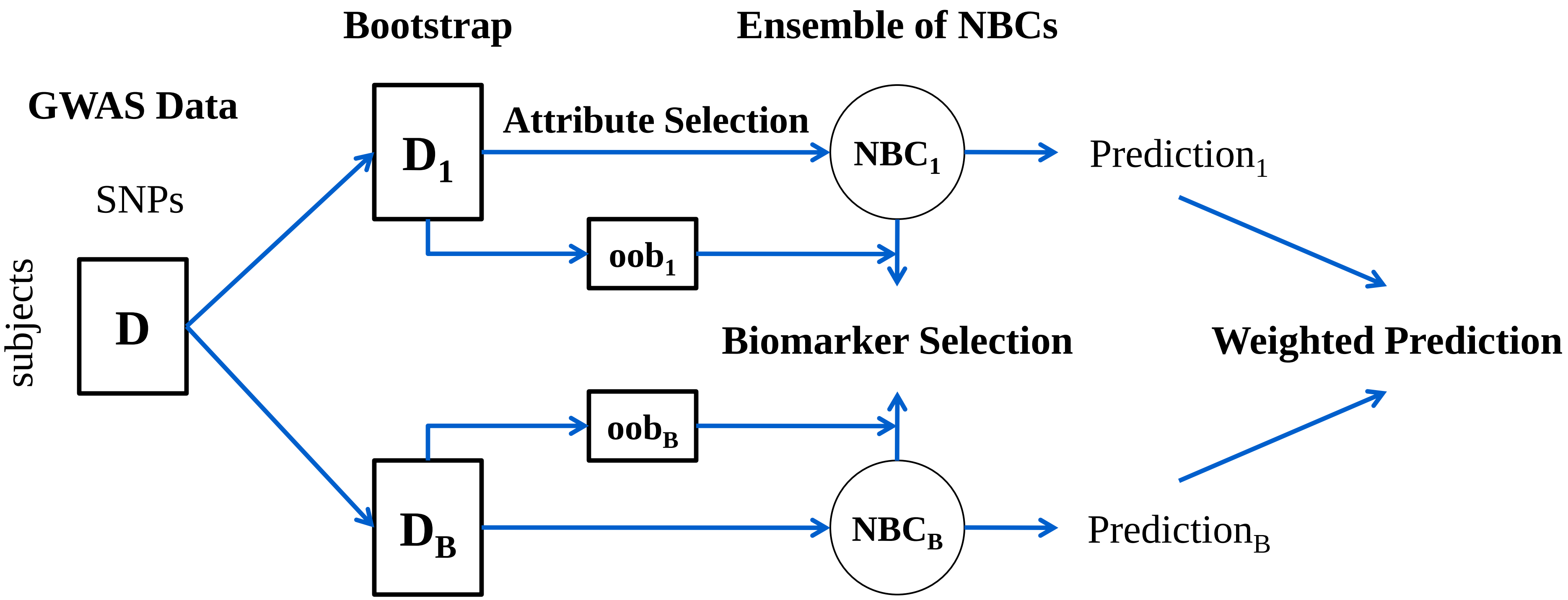
- B bootstrap replicates, sampled with replacement from D
- B Naive Bayes Classifiers, each trained on a D_b
- Outcome: average of the B predictions

NBC attribute selection (SNPs)



- Ranking: training error when SNP is used as single attribute
- Selection: top ranked, uncorrelated SNPs ($r^2 < 0.1$ if $dist < 1$ Mb)
- Number of selected attributes increased, as long as classification accuracy increases on the Out-Of-Bag (OOB) sets

Biomarker Selection



- Random permutation of the genotype of NBC attributes in OOBs
- Measure decrease in accuracy on OOBs
- Wilcoxon signed-rank test for significance

Results

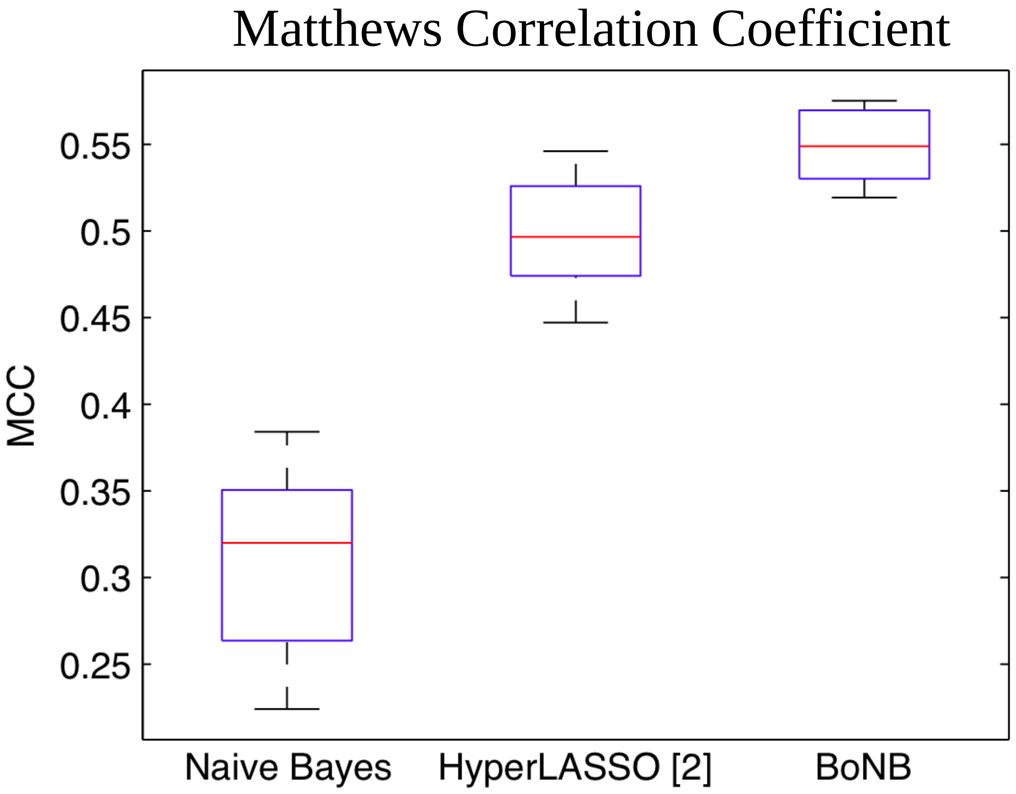
WTCCC case / control study on Type 1 Diabetes

- 458376 SNPs, 1963 T1D cases, 2938 controls

Biomarker Selection

rs ID	chr	gene
rs6679677	1	RSBN1
rs9273363	6	MHC region
rs3101942	6	MHC region
rs492899	6	MHC region
rs6936863	6	MHC region
rs805301	6	MHC region
rs9275418	6	MHC region
rs2856688	6	MHC region

Predictive accuracy



Conclusions

- BoNB effective for both classification and biomarker selection
- Advantages of bagging:
  - Higher generalization ability
  - Sound and principled procedure for biomarker selection
- Advantages of Naïve Bayes:
  - No pre-specified model of genetic effect
  - Seamless handling of missing values