BITS :: Call for Abstracts 2024 - Oral communication

Type	Oral communication
Session	Bioinformatics AI, Models and Tools
Title	Machine Learning applied to association studies on small and unbalanced sample data
All Authors	Dagnogo Dramane (1,5), Mirko Treccani(1), Cristina Patuzzo(1), Pietro Manuel Ferraro(2), Giovanni Gambaro(2), Dago Dougba Noel(3), Giovanni Malerba(4)

Affiliation

(1) Department of Neurosciences, Biomedicine and Movement Sciences, University of Verona, Italy.

(2) Division of Nephrology, Department of Medicine, University of Verona and Azienda Ospedaliera Universitaria Integrata, Verona, Italy

(3) Unité de Formation et de Recherche (UFR) des Sciences Biologiques, Département de Biochimie Génétique, Université Peleforo Gon Coulibaly, Korhogo, Côte d'Ivoire.

(4) Department of Surgery, Dentistry, Paediatrics and Gynaecology, University of Verona, taly

(5) Department of Cellular, Computational and Integrative Biology, University of Trento, Italy

Motivation

The performance of an association study depends on several aspects, including balanced design.A small sample size leads to underpowered statistical tests, increasing the risk of Type I (false positive) and Type II (false negative) errors. On the other hand, sample imbalance can bias results and limit the generalizability of the findings. Undersampling combined with ensemble machine learning (ML) could improve the predictive performance of the model, particularly in the case of unbalanced datasets, offering practical benefits for real-world data. This work aims to implement a machine learning (ML) model to enhance signal intensity and their detection from small sample sizes and unbalanced samples.

Methods

We simulated 4 datasets composed of 1,700 individuals and 2,100,598 SNPs using SeqSIMLA2 and using the reference variants' structure of the European individuals from 1000 Genome Project. We simulated data to mimicked real case/control data as much as possible, by setting prefixed SNPs with odds ratios ranging from 1.4 to 2.4 while the individual's number was set to 1,700. D1 consists of 800 Cases vs 900 Controls and 11 risk SNPs, D2 is composed of 800 Cases vs 900 Controls and the 5 risk SNPs showed significant after GWAS on D1 with a fix odd ratio of 1.9 for all the risk SNPs, D3 consist of 800 Cases vs 900 Controls and 12 risk SNPs, located on Chromosomes 1,2,3,6,10,18 and 19. D4 has the same parameters as D3 except for the Case/Control distribution which is 213 Cases and 1,487 Controls. These datasets underwent standard Quality-Control (QC) procedures including the removal of low-quality variants and individuals, and Hardy Weinberg Equilibrium following plink standard QC (--maf 0.01 --geno 0.01 --mind 0.02 --hwe 1.e-6), the association test has been performed using Plink2 generalized linear model using the parameter --glm. Features achieving the nominal significance (p-value < = 0.05) were selected, resulting in 52,214 SNPs for D3 and 45,055 SNPs for D4. We developed a Nested CatBoostClassifier (NCBC) model composed of a Random under sampler and a Catboostclassifier. The NCBC model has been trained on a subset of both datasets (80%) and tested on the remaining (20%). Evaluation of NCBC results was based on ROC AUC score, Accuracy score and the balanced accuracy score. Key features have been assessed using the built-in feature information gain (FIG) of the catboost module (FIG >= 0.1) and the Shaplet Additive Value (SHApvalue) of the module shap (shapvalue >= 0.01). Relevant SNPs of the imbalance data (D4) underwent a comparison analysis with the GWAS summary and finally, we assessed the impact of Shapvalue important SNPs on the sample distribution in the population using a sequential homogeneity test.

Results

All risks SNPs having a minor allele frequency >= 0.1 and odds ratio != 1 (+/- 0.2) (5 SNPs) showed significance in the association study of the dataset D1. Additionally, undefined significant associations consistently appeared in the same haplotype region of a defined risk allele. The risk SNPs simulated (snp914852, snp270912, snp947137, snp529979, snp1747883, snp1338407, snp146692, snp145838, and snp1793026) have a frequency ranging from 0.0197 to 0.46,73 and an odd ratio from 0.6539 to 3.3257. In D2, where all the risk SNPs have a maf > 0.1 and a fix odd ratio value of 1.9, all the risk SNPs were significant or in the peak of their respective Chromosome position. GWASs of D3 and D4 showed evidence of signal intensity amplitude increase in the balanced dataset with significant association peaks (p-value <= 5*10-8) on chromosomes 6, 10, and 19. The same peaks are observed in imbalance dataset (D4) with lower statistical support (p-value > 5*10-8) except for the peak on Chromosome 19. Simple catboost model highlighted comparable metrics in the balanced dataset (Accuracy: 67% accurate prediction, Balance accuracy: 66.9% accurate prediction, and ROC AUC score 71%), on the contrary in the imbalance dataset there is a large variability between them, 88.5% of accurate prediction when looking at the Accuracy, 51.2% for Balanced accuracy and ROC AUC score achieved a value of 51% . These results were confirmed by the confusion matrix, showing

that the prediction has been made on the majority class (1: Unaffected individuals) since this class accounts for 87.47% percent of the dataset (D4), predicting this class at 100% and making errors on 98% of the disease class, the accuracy remains higher. The NCBC applied to the datasets (D3 and D4), increased the prediction score of +27% (from 0.51 to 0.7 or a 50% of increase) (Figure D) in the imbalance data. Moreover, the prediction increased its value in both classes with a score of 82% in class 0 (Affected) and 74 % in class 1 (Unaffected) (Figure C). We also observed a slight improvement on the general prediction score of the balance dataset of +2.5% (figure 6B), the improvements are higher for the Affected class (class 0) with an increase of 7% (from 65% to 72%) (figure A).

Info	
-	
filename	-
Figure	
-	
Availability	https://drive.google.com/file/d/1Vxh1oCH4JSy8uPfXFqrGzsNMQ6zFtjav/view?usp=drive_link

Dissemination Material

Social

-

Summary

We implement a new machine learning model composed of a random undersampler and Catboostclassifier . This combination increased the accuracy of the prediction (from 51% to 78%) in imbalance data, it effect is more pronounce when we are interest in intra-category prediction where it improved the prediction score of Affected patients from 2% to 80%.

Corresponding Author

Name, Surname	Dramane, Dagnogo
Email	dramane.dagnogo@univr.it
Submitted on	02.05.2024