An application of Random Forest to personalized medicine.

S. Cabras[1], M. E. Castellanos[2] y N. Pirastu[3]

[1] Department of Mathematics, University of Cagliari.

[2] Departamento de Estadística e Investigación Operativa. Universidad Rey Juan Carlos.

[3] Shardna Life Science S.P.A. Pula

Scientists all over the world are looking for the genetic variants that underlie to common diseases. However most association techniques used for these studies consider only one genetic variant at a time and rarely the interactions between them. For this reason we looked ensemble methods applied to a known case: beta-thalassemia mutation carriers. Beta-thalassemia is a genetic disorder caused by a mutation inside the beta-hemoglobin gene. Only homozygous individuals for the mutation manifest the clinical traits of the disease, however carriers, although completely sane, show a reduced mean cell volume (MCV) of red blood cells, and this parameter can be used to identify them. In our study we show that we can trace back the position of the mutation out of about 500000 single nucleotide polimorfism (SNPs) genotyped in 500 samples that come from Talana a small village of Sardinia using Random Forest (RF) (Breiman, 2001). We were also able to determine the probability of a person to carry the beta-thalassemia mutation. The relatedness of the samples used and the knowledge of the presence of the same mutation all over Sardinia makes the beta-thalassemia mutation easier to find. Nonetheless the amount of markers measured on the genome (about 500 thousands) and the presence of other genes, which could affect MCV makes the finding not trivial. This situation many variables for few samples, that in a near future will rise increase to 1.8 million, is difficult to manage by classical regression methods or even Bayes methods such as the Bayesian Additive Regression Trees (BART) proposed by H. A. Chipman, E. I. George and R. E. McCulloch (2006). For this reason we propose to use RF in order to estimate the probability of MCV alteration given a certain genomic profile. Finally, another advantage of RF is that they are highly parallelizable and they are actually implemented for running in a cluster of CPUs which makes them suitable for studying such a high number of variables which require millions of iteractions.