8th International Conference of the Brazilian Association for Bioinformatics and Computational Biology
OCTOBER 14th to 17th
X - meeting 2012
Unicamp Convention Center | Campinas - São Paulo

An efficient implementation of Relief-F algorithm for Genome-wide association studies

Renato Shibata, Roberto Higa
Unicamp, Embrapa

Background: Genome-Wide Association Studies (GWAS) is an analysis of datasets containing individuals genotyped and Single Nucleotide Polymorphisms (SNP) besides a vector containing their corresponding phenotype.Number of individuals and SNP are both around hundreds of thousands. GWAS 's goal is finding out genomic regions implied in the phenotype by identifying associated SNPs with these regions. In Pattern Recognition, GWAS is a variable selection problem where SNPs correspond to independent variables and phenotype to dependent variables. According to the phenotype, the variable selection problem is modeled in a context of either classification or regression. Due to the fact that there are much more variables than samples, implementation must deal with statistical and computational restrictions. One way to circumvent these restrictions is using a two-step strategy.In the first step, it's possible to use a simple method to decrease the number of variables ,from hundred of thousands to just a few thousands.So the second step can use a computationally more demanding method. In this work we present an implementation of the algorithm Relief ,used in the first-step of a two-step strategy for GWAS. The family of algorithms Relief perform a heuristic procedure to score variables, aiming to select the top ranked ones, according to their relevance in discerning objects belonging to different classes. In comparison to other heuristic methods such as ID3, for instance, Relief is computationally more efficient and returns more accurate results. Although Relief can be adapted for regression context, in this work we approach only for classification context. Results: Our current implementation of Relief algorithm is written in C++ and approaches classification problems only. Genotypes datasets are represented in RAM memory using only two-bits by SNP so in this way a large number of SNPs can be managed in a fairly modest computer. In order to improve the algorithm computational performance as well as to select more representative instance , we used a combination of a Relief variation called Relief-F and data structure called KD-Tree. We compared our Relief-F+KD-Tree implementation results to those obtained by using the sofware Weka. Preliminar results showed that although both implementations presented equivalent results, our implementation is much more robust when dealing with larger datasets, which is the case of GWAS. Conclusions: We concluded that our implementation of Relief-F+KD-Tree is a promising approach for a fast, flexible and reliable first-step strategy approaching GWAS as a variable selection problem. Future works include to implement a Relief adaptation for regression problems as well as to test our implementation using real GWAS datasets.

Keywords: gwas, snp, relief-f, association studies.
Concentration area: Integration