

A Genetic Programming Algorithm for Association Studies

Robin Nunkesser

Department of Computer Science, TU Dortmund, 44221 Dortmund, Germany
Robin.Nunkesser@tu-dortmund.de

Abstract. The analysis of genetic association is useful for identifying genetic factors that may contribute to a medical condition. An important subarea are case-control studies on single nucleotide polymorphism (SNP) data, i.e. data on genetic variations that occur when different base alternatives exist at a single base pair position. The major goal of these studies is to identify SNPs and SNP interactions that lead to a higher disease risk.

We present a Genetic Programming algorithm called GPAS (Genetic Programming for Association Studies) for case-control association studies which outperforms other regression and discrimination approaches on real and simulated SNP data examples. On these examples, GPAS is also able to identify high-order interactions of SNPs with a high odds ratio, which are not found by other feature selection methods. The algorithm is implemented in an extendible R package called RFreak, which also allows an easy modular implementation of further Genetic Programming and Evolutionary Computation algorithms.

Finally, we discuss more areas the algorithm is applicable to and extensions enabling GPAS to work on different types of association studies.

Keywords: Association studies, Genetic Programming, Classification, R

References

- HOH, J. and OTT, J. (2003): Mathematical multi-locus approaches to localizing complex human trait genes. *Nat. Rev. Genet.* 4, 701–709.
- NUNKESSER, R., BERNHOLT, T., SCHWENDER, H., ICKSTADT, K. and WEGENER, I. (2007): Detecting high-order interactions of single nucleotide polymorphisms using genetic programming. *Bioinformatics* 23 (24), 3280–3288.
- RUCZINSKI, I., KOOPERBERG, C. and LEBLANC, M (2003): Logic regression. *J. Comput. Graph. Stat.* 12, 475–511.