

MCMC or Genetic Programming?

Comparing Search Algorithms for the Analysis of Genotype Data

K. Ickstadt, T. Bernholt, A. Fritsch, R. Nunkesser, H. Schwender, and R. Fried

Collaborative Research Center SFB 475, University of Dortmund, Germany

ickstadt@statistik.uni-dortmund.de, fritsch@statistik.uni-dortmund.de, fried@statistik.uni-dortmund.de

Background

Single Nucleotide Polymorphism (SNP)

- Most common type of variations in human DNA which occurs when different alternatives exist at a single base pair position.
- Diploid genome \Leftrightarrow Each SNP can take three different genotypes:
 1. Both bases are the more frequent variant (*homozygous reference*).
 2. One is the more, the other the less frequent variant (*heterozygous variant*).
 3. Both bases are the less frequent variant (*homozygous variant*).

Goals in Case-Control Studies

- Identification of SNP interactions explanatory for the case-control status.
- Construction of a classification rule based on SNPs and SNP interactions.

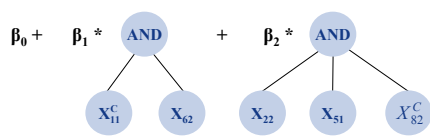
Logic Regression

- Predicts the case-control status based on Boolean combinations of binary variables.
- Logic trees are used for both the representation of logic expressions and for the generation of new trees in the search for (the set of) the best model(s).
- Either a single tree can be grown or multiple trees can be adaptively grown and combined by a logistic regression.
- Currently implemented search algorithms: Simulated Annealing (*LR*), (a greedy search) and an MCMC based algorithm (*MCLR*).

New Search Strategies

- Full Bayesian MCMC approach (*FBLR*).
- Genetic Programming (*GP*).

Fully Bayesian Logic Regression



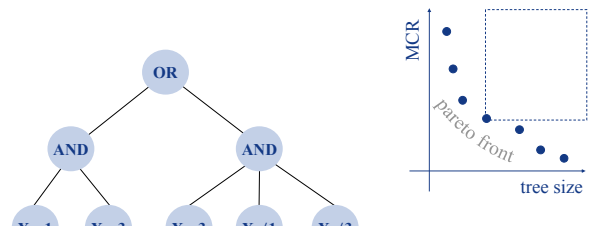
1. Start with the null model.
2. Propose a new model according to the move set of FBLR.
3. Accept the model with a certain probability based on the Bayes factor.
4. After initial burn-in keep a sample from the posterior distribution of models.

Advantages:

- Priors on the β s.
 - No penalty parameter on model size.
- } Inclusion of genetic information such as pathway information.

FBLR	Move	GP
	Mutation:	
x	Add a leaf / subtree	x
x	Remove a leaf / subtree	x
x	Change a leaf	x
	Crossover:	
	Interchange subtrees	x
	Merge leafs of two subtrees	x

Genetic Programming



1. Start with a randomly selected tree.
2. Generate new trees using the moves of GP. The moves generate trees in disjunctive form.
3. Select the pareto-optimal trees with respect to tree size and MCR.

Advantages:

- Mimics genetic processes, uses Crossover to reach better solutions.
- Computes a pareto front \Leftrightarrow A solution for every tree size.
- Multivalued logic and disjunctive form \Leftrightarrow AND-tree = SNP interaction.

Application to Simulated SNP Data

To compare the new search algorithms for logic regression with the already existing ones, we simulate SNP data and the case-control status as follows:

- For each of 50 SNPs, the values of 1,000 observations are drawn from a Bin(2, 0.25) distribution.
- For each observation, the case-control status y is drawn from a Bernoulli distribution with mean $\text{Prob}(Y=1)$, where

$$\text{logit}(\text{Prob}(Y=1)) = \beta_0 + \sum_{i=1}^3 \beta_i L_i$$

with $L_1 = \{S_6 \neq 1\} \wedge \{S_7 = 1\}$, $L_2 = \{S_8 = 3\}$ and $L_3 = \{S_3 \neq 1\} \wedge (\{S_9 \neq 1\} \vee \{S_{10} \neq 1\})$.

To investigate differing effect sizes, we consider three models with differing values for $\beta_1 = \beta_2 = \beta_3$. While $\beta_0 = -0.2$ in all three models, β_1 is either set to 1, 1.5 or 2.

For each of these three settings, two data sets – one training set and one test set – are generated.

Each of the search algorithms is applied to the training set to build a model which in turn is used to compute the misclassification rate (*MCR*) on the test set.

	$\beta_1 = 1$		$\beta_1 = 1.5$		$\beta_1 = 2$	
	Iteration	MCR	Iteration	MCR	Iteration	MCR
True Model	–	0.355	–	0.305	–	0.274
LR	250,000	0.393	250,000	0.319	250,000	0.347
MCLR (Single)	125,000	0.358	125,000	0.365	125,000	0.298
GP	50,000	0.365	50,000	0.329	50,000	0.307
MCLR (Multiple)	125,000	0.358	125,000	0.310	125,000	0.274
FBLR	125,000	0.361	125,000	0.322	125,000	0.288

