Feature selection for high dimensional regression using local search and statistical criteria

Julie Hamon, Clarisse Dhaenens, Gaël Even, Julien Jacques

Feature selection for high dimensional regression using local search and statistical criteria

J. Hamon[1,2,3], C. Dhaenens[2,3], G.Even[1], and J. Jacques[3,4]

[1] Gènes Diffusion Douai
julie.hamon@inria.fr, g.even@genesdiffusion.com
[2] LIFL, UMR CNRS 8022, Université Lille 1
clarisse.dhaenens@lifl.fr
[3] INRIA Lille-Nord Europe
[4] Laboratoire Painlevé, UMR CNRS 8524, Université Lille 1
julien.jacques@inria.fr

1 Motivation

Genomic selection is a genetic evaluation of animals from their DNA (extracted using biological samples such as blood or hairs, or biopsy), based on a huge number of markers covering the whole genome. The basic principle was established by Meuwissen, Hayes and Goddard in 2001 [4]. In this context, an important objective of genomic selection consists in searching explicative markers for a phenotype (quantitative trait characterising an animal) under study.
Nowadays, with the development of new technologies such as high-throughput genotyping and sequencing, it is possible to conduct such studies and read genomic information on around 800,000 markers on more and more subjects.
Recently, these datasets are studied to establish predictive models using genomic information. However, in addition to biological constraints (such as sample storage, time consuming and costly experiments, ...) data analysis needs to be improved.
Lots of significative markers identification studies have been done for qualitative traits, often binaries (disease or not). Here we deal with quantitative traits such as milk production or meat quality, so the challenge is to find a predictive model allowing to select the best animals for these phenotypes.

2 A data mining problem

This context of genomic selection in animal studies involves a lot of characteristics such as familial relationships between animals or correlations between markers. We will first leave out these characteristics to focus on the underlying datamining problem in the case where the number of animals (n) is less than the number of markers (p).
With the increase of the number of markers, sequential methods (markers analysed one by one with a linear regression) becomes unsuitable, time consuming and do not take into acount possible interactions between markers. Among the classical methods used to deal with a large number of features, we can cite LASSO [6], PLS [3] or BLUP (model specific to animal genetic) [2]. However, to have a predictive model easily interpretable and comprehensible, we can jointly select pertinent features (markers) and compute the explicative model. We propose to modelize the problem as follows:

$$Y_i = \beta_0 + \sum_{j=1}^{p} (\beta_j z_j X_{ij}) + \epsilon_i \ ,$$

with Y the trait under study ($Y \in \mathbb{R}^n$), X_j the markers studied ($X_{ij} \in \{-1, 0, 1\}$), ϵ_i i.i.d $\sim N(0, \sigma^2)$, and $z_j = 1$ if the marker j is selected, 0 otherwise. The objective is to estimate the β_j and z_j. As z is a discrete vector, in a finite set $\{0, 1\}^p$ we cannot estimate it when p (the number of features) is large. Determining z_j values is equivalent to determine features that are selected and participate to the regression model. This problem is a typical feature selection problem known in data mining. The particularity here is to address regression and not classification task with the selected features.

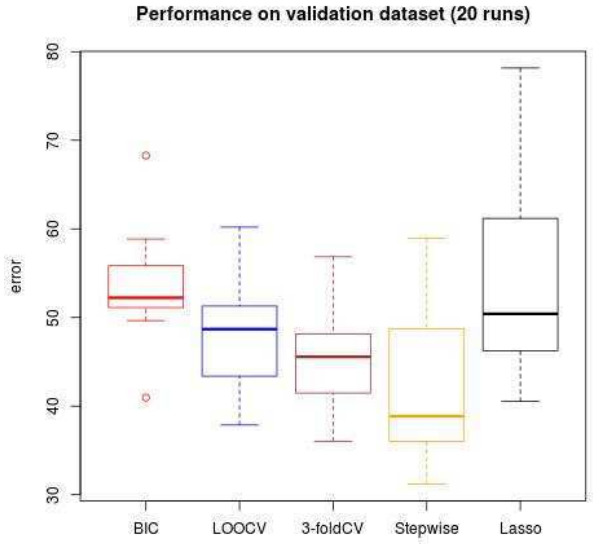
3 Approach: ILS

Feature selection is a combinatorial problem that may be addressed by combinatorial optimization methods [5]. In particular, when the number of features is large, metaheuristics could be used.
In our specific context of regression, we propose to combine an iterated local search (ILS) with a statistical evaluation of a multivariate regression. Indeed, a particular attention should be paid on the way to evaluate the regression model quality. Therefore, we will compare several approaches.
The local search works on subsets of features (binary vectors indicating whether the feature is selected or not), and explores the neighborhood of the solutions (adding or removing a feature) in order to obtain a reduced number of features. When a local optimum is reached, the current solution is perturbed (removing several features) to continue the search. The search stops when a given number of iterations without improving the best solution is reached. As several ways exist to evaluate the quality of the regression model, we compared three criteria in order to analyse their impact on the performance of the local search. These three statistical criteria are: a bayesian information criterion (BIC), a k-fold cross validation (k-foldCV, here we choose k=3) and a leave-one-out cross validation (LOOCV).

$$z_j \xrightarrow{Regression} \text{likelihood} \xrightarrow{BIC} \text{fitness}$$
$$z_j \xrightarrow{k-foldCV(Regression)} \text{fitness}$$
$$z_j \xrightarrow{LOOCV(Regression)} \text{fitness}$$

4 Validation

In order to assess the quality of the proposed method and to compare the three criteria, simulated data have been generated. A training dataset (with $n \ll p$) and a validation dataset are used. Moreover, in order to make a comparison with the literature, we ran two classical statistical approaches generally used to predict a quantitative trait from a large number of features [1]: a stepwise regression method and a penalized regression method (LASSO [6]).
Results shows that our approach generates interesting explicative models. We can remark that the stepwise method manages to obtain better results. This is explained by the overfitting (encountered as $n \ll p$) that stepwise avoids by selecting less features using a threshold to decide whether adding a feature or not. One of our perspective is to introduce such a threshold in our approach.



References

1. T. Hastie, R. Tibshirani, J. Friedman. The Elements of Statistical Learning, 2009.
2. C. R. Henderson. Best Linear Unbiased Estimation and Prediction under a Selection Model. *Biometrics*, 31(2):423-447, 1975.
3. N. Long, D. Gianola, G.J.M Rosa, and K. A Weigel. Dimension reduction and variable selection for genomic selection: application to predicting milk yield in holsteins. *Journal of Animal Breeding and Genetics*, 128(4):247-257, August 2011.
4. Meuwissen, T. H. E., Hayes, B. J. and Goddard, M. E. . Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. *Genetics Society of America*, 2001.
5. Saeys, Y., Inza, I. et Larranaga, P. A review of feature selection techniques in bioinformatics, *Bioinformatics*, vol. 23, n. 19, p. 2507-2517, oct. 2007.
6. T. T. Wu, Y. F. Chen, T. Hastie, E. Sobel and K. Lange. Genome-wide association analysis by Lasso penalized logistic regression. *Bioinformatics*, 25(6), 2009.