

An application of Random Forest and Hungarian Method to genome wide association of asthma in a genetic isolate of Ogliastra¹

S. Cabras^{2(*)}, N. Pirastu^(*,3), L. Casula⁽⁴⁾, M.E. Castellanos⁽⁵⁾, I. Persico⁽⁴⁾, A. Sassu⁽²⁾, G. Biino⁽⁴⁾, S.R. Del Giacco⁽⁶⁾, M. Pirastu⁽⁴⁾
²Department of Mathematics, University of Cagliari (Italy), e-mail: s.cabras@unica.it

Abstract: We consider the problem of predicting the asthma status by estimating epistasis, namely, interactions among genetic variants and environment factors with asthma. We estimate epistasis using Random Forest (RF) using a data set with 500000 genetic variants and environmental variables that have been measured on about 200 cases and controls. The training sample for the RF is selected among a large database that contains also the genealogy tree that able us to relate all subjects. With the Hungarian method we choose the most inbred controls related to the cases. The exogenous genetic variability of such training sample is reduced if compared with that in open populations of unknown kinship. This allows us to successfully build a RF classifier which has an acceptable prediction error of the asthma status. Moreover, one of the most important genetic variant, that we associate with the asthma, has been also reported to be functionally associated with asthma in other studies.

Keywords: Asthma, Ensemble methods, Epistasis, Genetic isolate, Genome wide association, Hungarian method

1. Introduction

In this paper we address the problem of estimating DNA markers related to a complex disease. This problem is dealt with a large sets of DNA, say 500000 Single-Nucleotide Polymorphisms (SNPs) observed on highly inbred population belonging to a genetic isolate. A further challenge of this analysis is that we know and must use the relativeness degree between cases and potential controls through an available genealogy tree. Finally, for most of the individuals is available a set of covariates that reflects individual habits which, in principle, may be relevant for the studied disease if they interact with SNPs. We will refer to such covariates as environmental variables. Complex diseases make huge the space of possible statistical sampling models that relate SNPs and environmental variables to the disease, because epistasis involves interactions of unknown order. For instance, three or more genes may interact with one or more environmental variables in order to be a protective or a risk factor for the disease. Such interactions may be difficult to interpret because of the weak knowledge of gene functions and our interests mainly

¹The present paper is financially supported by the University of Cagliari and MIUR.

²Address of correspondence: via Ospedale, 72 - 09124 Cagliari

(*)These authors contributed equally to this work. ⁽³⁾Shardna S.p.A., ⁽⁴⁾Istituto di Genetica delle Popolazioni CNR, ⁽⁵⁾Departamento de Estadística y Investigación Operativa Universidad Rey Juan Carlos (Madrid),

⁽⁶⁾Dipartimento di Scienze Mediche Internistiche - Università di Cagliari

lies in a model with satisfactory prediction properties, measured by the prediction error. The situation of highly dependent observations may, at a first look, further complicates the analysis because one cannot assume usual parametric models for independent observations. However, knowing the kinship between individuals may allow us to obtain a suitable experimental design pairing controls to the most related cases. This design reduces the amount of genetic variability exogenous to the disease. Although the above mentioned framework is quite general and the statistical methods here discussed are of general applicability, we will look at epistasis in order to predict the asthma in a genetic isolate of Ogliastra in the Sardinian island (Italy). Data have been provided by Shardna S.p.A. (www.shardna.com). We dispose of a total of $I = 57$ asthmatic cases and $J = 151$ non asthmatics potential controls. Moreover, we dispose of a further sample of 1043 genotyped people which can be used for further validation analysis. The total amount of considered SNPs is $q = 429373$ each with two or three configurations. Most of them are out of genes and it is known that two adjacent SNPs are highly correlated with correlation decreasing with the increasing of the distance in the DNA. For each individual we know sex, age and habits on smoking, physical and sport activities. These covariates may potentially be related to the asthma.

2. Method

Our approach is characterized by the selection of the sample used to train the prediction model and its construction. Both are described in the next two subsections.

2.1 Sample selection

The prediction model we proposed to use need to be trained with a sample of cases and controls where we known the asthma status, SNPs configurations and habits. This is the training sample (TS). With TS we derive genetics variants and habits related to the asthma. Once the predictive model has been trained, we can use it to predict the asthma state on further individuals out of the TS. In order to successfully estimate epistasis we better reduce the exogenous genetic variability by coupling controls with the most related cases. For instance, if we consider two individuals: one asthmatic and the other not, we are more likely to efficiently estimate epistasis if they were brothers rather than two unrelated people. In order to do this we first summarize the relativeness among cases and potential controls by the *kinship index* deduced using the genealogy tree. The kinship index decreases according to the number of meiotic steps that separate two individuals. Let k_{ij} be the kinship index between case i , $i = 1, \dots, I$ and control j , $j = 1, \dots, J$, for each case i we have to chose the control j_i^* such to maximize $\sum_{i=1}^I k_{ij_i^*}$. This is an assignment problem solved with Hungarian method proposed by H. Kuhn and described, for instance in, Papadimitriou and Steiglitz (1982). The assignment matrix is showed in Figure 1(a) where bright rectangles indicate highly related individuals that form the TS. In order to appreciate the quality of the obtained TS we perform Fisher exact tests between the asthma and the q SNPs. It results that because of the particular selected sample, we have the same SNPs configurations between cases and controls, which produces p -values equal to 1 as showed in Figure 1(b). These SNPs are likely to be not related to the asthma but their variability is exogenous to the analyzed disease. However, Fisher exact tests only approximate epistasis up to first order interactions between asthma and single SNP

configuration, while our goal is to model epistasis with higher order interactions.

2.2 The prediction model

In order to obtain a prediction model we use Random Forest (RF) described in Breiman (2001). At the moment, this model has rarely been used in literature on genomic wide association studies, see for instance Díaz-Uriarte and de Andrés (2006) and de Edelenyi *et al.* (2008). RF is an ensemble of m non-parametric classifiers, in this case classification trees Breiman *et al.* (1984) with Gini loss function. RF allows to make prediction on asthma status Y conditionally on a certain profile $\mathbf{x}$ which represents SNPs and environmental variables. RF consists in draws of subsamples of size $n' < n = 57$ from the TS (the so called *in bag* sample) and a corresponding draw of $q' = \sqrt{q} < q$ variables used to build a tree. This is repeated $m \approx 10000$ times in order to have a forest built with the m random trees (hence the name RF). Let $Y = 1$ indicates asthma presence and $Y = 0$ otherwise, our parameter of interest is $\Pr(Y = 1|\mathbf{x})$ which is estimated using the proportion of the m trees where the profile $\mathbf{x}$ belongs to a node classified as $Y = 1$. We classify an individual as asthmatic if the estimated $\Pr(Y = 1|\mathbf{x})$ is greater than 0.5. With such approach we can make variable selection and prediction at the same time because trees use only subsets of $\mathbf{x}$ that reduce the prediction error in estimating Y with respect to a naive classifier that systematically classifies an individual as asthmatic. This classifier, for a balanced TS, has an initial error of 50%. Therefore, such subset of predictors is also likely to be associated with the asthma. It is worth mention that the prediction error is always evaluated on the set of non in bag sample, namely the *Out Of Bag* sample, which provides a more realistic evaluation of the prediction error than the one based on the in bag sample. As classification trees are able to model the interactions among subsets of $\mathbf{x}$ in predicting asthma then they also model epistasis and the final epistasis estimation corresponds to the estimated RF.

2.3 The analysis

Assuming that there are very few mutations of the DNA that are causing the asthma, we may find a lot of SNPs correlated and associated with the asthma just because they are near the mutation point. Because of the *Occam razor*, we want to isolate only the most associated SNPs and so we initially consider to run RF with $q' = q$, usually referred as *Bagging* procedure. This makes compete all q variable in predicting Y , thus eliminating variable collinearity due to SNPs correlations. We repeated RF eliminating the less important SNPs ending up with a list of 200 SNPs order from the most to the least important.

3. Results

The available environmental variables are less important in predicting the asthma than SNPs. From Figure 2(a) we can see how the prediction error decreases when using RF with the first q' important SNPs, in particular when $q' \approx 100$ the prediction error is about 15%, which is more than three times less than the initial error. We further validate our prediction model on the unvisited 1043 individuals. We obtain an association of the predicted asthma with the degree of physical activity, shown in Figure 2(b). In fact, individuals supposed to be healthy used to have moderate or intense physical activity,

while those classified as asthmatic have a scarce physical activity. Finally, the third most important SNPs is inside a gene associated functionally to non allergic asthma in rats (de Edelenyi *et al.*, 2008).

Figure 1: *Features of the training sample.*

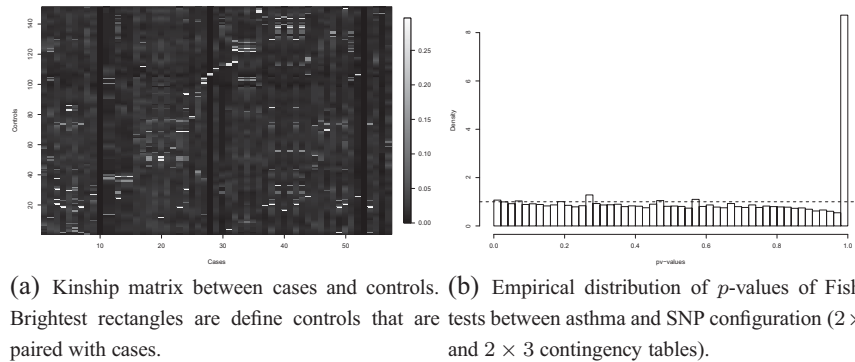
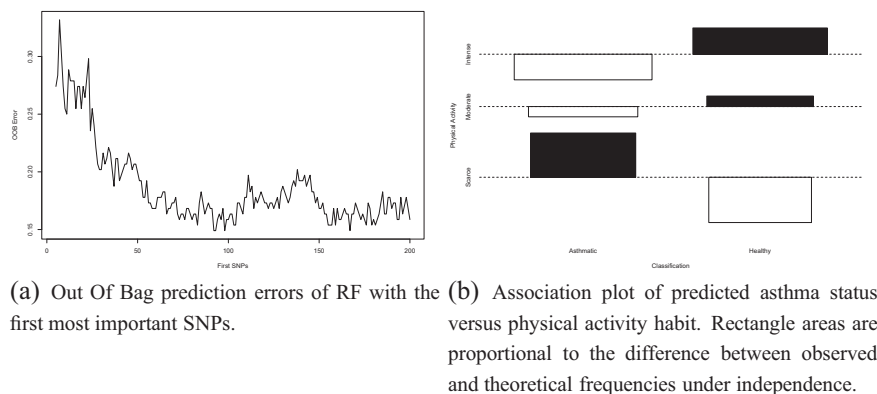


Figure 2: *Performance of the predictive model.*



References

- Breiman L. (2001) Random forests, *Machine Learning*, 45, 5–32.
- Breiman L., Friedman J., Olshen R. and Stone C. (1984) *Classification and Regression Trees*, Chapman and Hall/CRC Press.
- de Edelenyi F.S., Goumidi L., Bertrais S., Phillips C., Macmanus R., Roche H., Planells R. and Lairon D. (2008) Prediction of the metabolic syndrome status based on dietary and genetic parameters, using random forest., *Genes Nutr.*, 3, 173–6.
- Díaz-Uriarte R. and de Andrés S.A. (2006) Gene selection and classification of microarray data using random forest, *BMC Bioinformatics*, 7, 3, 11–22.
- Papadimitriou C. and Steiglitz K. (1982) *Combinatorial Optimization: Algorithms and Complexity*, Englewood Cliffs: Prentice Hall.