Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

Bayesian nonparametric clustering and association studies for candidate SNP observations

Charlotte Wang [a], Fabrizio Ruggeri [b], Chuhsing K. Hsiao [c], Raffaele Argiento [b,d,*]

[a] Department of Mathematics, Tamkang University, Tamsui District, New Taipei City 25137, Taiwan
[b] CNR-IMATI, Milano 20133, Italy
[c] Bioinformatics and Biostatistics Core, Division of Genomic Medicine, Research Center for Medical Excellence, National Taiwan University, Taipei 100, Taiwan
[d] School of Mathematics, Statistics and Actuarial Science, University of Kent, Canterbury CT2 7NF, UK

ARTICLE INFO

Article history:
Received 16 October 2015
Received in revised form 16 July 2016
Accepted 30 July 2016
Available online xxxx

Keywords:
Bayesian
Clustering
Dirichlet process mixture model
Exchangeable partition probability function
Genetic score
GWAS
Logistic regression
Nonparametric
Random partitions

ABSTRACT

Clustering is often considered as the first step in the analysis when dealing with an enormous amount of Single Nucleotide Polymorphism (SNP) genotype data. The lack of biological information could affect the outcome of such procedure. Even if a clustering procedure has been selected and performed, the impact of its uncertainty on the subsequent association analysis is rarely assessed. In this research we propose first a model to cluster SNPs data, then we assess the association between the cluster and a disease. In particular, we adopt a Dirichlet process mixture model with the advantages, with respect to the usual clustering methods, that the number of clusters needs not to be known and fixed in advance and the variation in the assignment of SNPs to clusters can be accounted. In addition, once a clustering of SNPs is obtained, we design an individualized genetic score quantifying the SNP composition in each cluster for every subject, so that we can set up a generalized linear model for association analysis able to incorporate the information from a large-scale SNP dataset, and yet with a much smaller number of explanatory variables. The inference on cluster allocation, the strength of association of each cluster (the collective effect on SNPs in the same cluster), and the susceptibility of each SNP are based on posterior samples from Markov chain Monte Carlo methods and the Binder loss information. We exemplify this Bayesian nonparametric strategy in a genome-wide association study of Crohn's disease in a case-control setting.

1. Introduction

Researchers nowadays prefer to test the association between multiple markers and a disease of interest in genetic association studies because the tests with multiple markers are more powerful, efficient, and biologically meaningful than single marker tests. Many statistical methods have been proposed based on those considerations, such as regularized regression models like lasso or ridge regression [1–5], gene-set enrichment analysis [4,6–8], pathway [9], and network analysis [10]. Those methods are helpful to analyze large-scale markers and their corresponding interactions in the same pathway or network, when the analytic genomic region is pre-defined. Such tools, however, may be limited when utilized on regions containing a great amount of genetic markers or at the genome-wide scale. When analyzing data with such size, there may

[*] Corresponding author at: CNR-IMATI, Milano 20133, Italy.
E-mail address: raffaele@mi.imati.cnr.it (R. Argiento).

http://dx.doi.org/10.1016/j.ijar.2016.07.014
0888-613X/© 2016 Elsevier Inc. All rights reserved.

be no complete information about the role of each gene and the interaction among them, so that figuring out the association between these markers and disease phenotypes can be challenging. Therefore, an important issue for scientists is how to cluster or categorize the genomic markers in advance, so that the dimension of the data can be reduced and the genetic markers are represented with several relatively small and manageable sets.

Most current clustering algorithms evaluate first the distance between objects and then group them according to certain criteria. The definition of distance can vary from Euclidean measure for continuous observations to counting measure for discrete data. The choice depends on the problem and also on the data characteristics. For discrete observations like SNP genotypes, similarity or dissimilarity measures can be employed. A common measure with a natural biological interpretation is the linkage disequilibrium, where allele frequencies per locus and haplotype phase need to be derived *a priori*, based on genotype data. This derivation involves the uncertainty in haplotype configuration, introducing even more parameters in either case-control or pedigree studies [11,12]. Other algorithms use mathematical formulations of similarity between SNP genotypes, including principle component analysis [13], k-means, and Hamming distance metric [14]. These tools are flexible in the sense that no biological information is required in advance. In most clustering algorithms, however, the decision on the number of clusters is a difficult task. Its choice as a stochastic parameter usually complicates modeling and increases the computational burden [15–17].

Clustering or partitioning can be easily dealt within a Bayesian nonparametrics framework through the Dirichlet process mixture models, which allocate data to clusters and determine their number [18,19]. Previous Bayesian applications in association studies either assumed two fixed clusters, i.e., associated vs. non-associated genes, and used Bayes factors for hypothesis testing, or applied a mixture model for every single marker [20–22]. No clustering procedure or multiple-marker effects were considered, and markers were examined individually, assuming exchangeability of their parameters.

From the modeling point of view, Dirichlet process mixture (DPM) models do not require the specification of the number of mixture components and the clustering procedure can be viewed as a Chinese restaurant process (see [23,24] for more details). Inference on the number of clusters and mixture model parameters estimation are unified and performed by a suitable Markov chain Monte Carlo (MCMC) algorithm, also integrating out the nonparametric component by a so called Polya urns Gibbs scheme (see, for instance the research by Neal [25]). For more details on model based cluster analysis in Bayesian nonparametric setting we refer to [26,27].

Bayesian models for cluster analysis are becoming more and more popular even in the genetic epidemiological and biomedical literature. Among the other papers, DPM models with Gaussian kernels are used to cluster microarray gene expression data [28–30]. Our approach differs from the previous papers since SNP genotypes take only three possible values and thus we consider a multinomial mixture model [31,32]. It is worth mentioning that our goal is very similar to the one in [31], although with a different approach. They clustered individuals in groups (e.g., high risk, average risk and low risk for a certain disease) and then identified the covariates which were influent in clustering with DPM. In our approach the procedure is reversed, since we first cluster the SNPs according to a DPM model with multinomial kernels, and then we investigate which groups of SNPs affect the disease risk of an individual. In [33] we presented a model similar to the one discussed in this paper, by considering a wide class of processes, namely the normalized generalized gamma processes (NGG), as mixing distribution in the hierarchical mixture model. We stress here that in the current paper the focus is on the application, i.e., the association study between groups of SNPs and a disease, while in the previous work we were more interested in proving the feasibility of normalized generalized mixture model in addressing real problems, in modeling and, furthermore, in providing a review of the model and its current applications.

More in detail, in the current work SNP genotypes are categorical and thus the codings do not affect the inference. Following the allocation of SNPs in clusters, we compute the genetic score of each cluster to investigate the cluster effects under the generalized linear mixed effect model (GLMM). The risk of each cluster can be evaluated based on its corresponding posterior probability and, in addition, the effect of each single marker inside the cluster can be evaluated as the mean of a suitable posterior functional.

The rest of the paper is organized as follows. In Section 2 we present a Bayesian nonparametric approach which clusters SNPs based on the observed numbers of counts of minor alleles via a Dirichlet process mixture model. We also give some detail on the Gibbs sampler to perform posterior inference and to compute the posterior clustering based on the so called Binder loss information. In Section 3 we propose a genetic score to investigate the cluster effect through a link function in GLMM. Each cluster can be identified to be positively or negatively associated with the disease phenotype based on its corresponding posterior probabilities and single SNP effects. In Section 4 we apply the analyses to a study of Crohn's disease from Wellcome Trust Case Control Consortium [34]. Results from the proposed method are compared with other analyses to evaluate the performance. Concluding remarks are given in Section 5.

2. Bayesian nonparametric model-based clustering algorithm

The SNP data we are going to consider belong to M different chromosome regions, and in this work we are going to suppose that clustering of SNPs are independent across different regions; it is well recognized that different, non-adjacent, chromosome region may not be passed together from parents to offspring due to the so called random crossover, so that independence among different regions may be assumed. From a modeling point of view, we are then going to fit M independent Dirichlet mixture models, one for each region.