Support vector machine adaptation to biallelic SNP data

Ricardo Cao[1], Wenceslao Gonzalez-Manteiga[2] and Manuel Garcia-Magariños[2]

[1] Department of Mathematics, University of A Coruña, Spain
[2] Department of Statistics and Operations Research, University of Santiago de Compostela, Spain

The past decade has witnessed the appearance of large amounts of genetic data, due to the association of both computational and molecular technological developments. As a result, case-control association studies interested in disentangling the genetic causes of complex diseases have increased exponentially. Nevertheless, most of the recent success in the field of statistical genetics has come from identifying genes with substantial non-interactive effects. Common heritable diseases (sporadic breast cancer, schizophrenia, diabetes,...) do not generally show Mendelian models of transmission, and are thought to be under the influence of multiple, and possible interacting, genes. At this point, there is a crucial need for massive efforts in development of new highly qualified methodologies to unravel the complex genetic basis of non-mendelian disorders (Thornton-Wells *et al*, 2004).
The support vector machine (SVM) is a kernel-based learning system (Vapnik, 1995), usually applied for solving problems in nonlinear classification. The idea is to construct an optimal separating hyperplane in a high-dimensional dot product space. Its logic is based in the fact that data separation can be made much simpler increasing the dimension, even up to infinity. In spite of this, SVMs avoid overfitting and are computationally feasible, although they give ocasionally rise to a strong computational burden. SVMs have been already used in many scientific fields, almost always successfully.
In this work we present a Support Vector Machine (SVM) adaptation to biallelic Single Nucleotide Polymorphisms (SNP) data. The new method consists of adapting SVM to the type of data we are working with. Most of the kernels frequently used in the scientific literature were specifically developed to be applied on continuous data while biallelic SNPs are categorical variables with three possible (coded) values. Any kernel arises as a similarity measure that can be thought of as a dot product in a so-called feature space. In our case, it is fundamental to turn the categorical (coded) SNP values into the binary values of their corresponding pair of alleles, since it is not natural to define a similarity measure directly on coded ternary data. The new kernel computes its value as the weighted sum of the importance measures from those alleles and those pair of alleles in which the two individuals coincide. Importance measures are obtained from allele frequencies in the sample. Considering pairs of alleles in the kernel allows us to look for SNP-SNP interactions responsible for disease.
Simulated biallelic SNP data obtained with SNaP software were used to evaluate the ability of our SVM adaptation both to discover causal SNP interactions associated with disease and to reduce case-control prediction error. GRID computational tools available at the Galician Supercomputing Centre (CESGA) helped to reduce significantly computation time.

References

Vapnik V.N. (1995) *The Nature of Statistical Learning Theory*. Springer-Verlag, New York.

Thornton-Wells T.A., Moore J.H. and Haines J.L. (2004). Genetics, statistics and human disease: analytical retooling for complexity. *Trends in Genetics*, **20**, 640–647.