A new similarity measure for the analysis of genetic data in complex phenotypes

University of Calabria

Alberto Montesanto[1], Vincenzo Lagani[2], Fausta Di Cianni[2], Domenico Conforti[3] and Giuseppe Passarino[1]

1. Department of Cell Biology, University of Calabria, Rende, Italy
2. Supercomputing Center for the Computational Engineering - CESIC, NEC Italia, Rende, Italy
3. Department of Electronics, Informatics and Systems - DEIS, University of Calabria, Rende, Italy

Introduction

Recent technological advances have led to the accumulation of a remarkable bulk of genome wide data on genetic polymorphisms. Consequently, there has been a growing interest for the possibility to carry out genomewide association studies for a variety of human complex traits, including cancer, cardiovascular diseases, aging and so on. However, the development of new statistical and informatic tools for the effective processing of these data has not been equally fast. Recently, kernel-based methods have attracted attention of many researchers in the field of statistical data mining and knowledge discovery. In fact, many kernel-based methods have been developed, e.g. kernel principal component analysis (KPCA), kernel logistic regression (KLR), kernel Fisher discriminant analysis (KFDA). They have become popular tools for classification, clustering, and regression analysis in the machine learning community since the introduction of support vector machines (SVMs) during the early 1990s.

Support vector machines represent a set of supervised learning methods that, taking advantage from a kernel function, are used for classification problems. They are widely used to analyze high-dimensional gene-expression data. However, the performance of a SVM classifier is strongly related to the similarity measure used into the classifier (kernel function). In fact, the choice of this function significantly influences the results of the kernel-based algorithms. The influence varies with data sets and the types of applications. In spite of their well known statistical power, to date, Machine Learning literature counts relatively few examples of works focused on the development and application of data mining methods specifically devised for the analysis of genetic polymorphisms.

Aim of our work was to define a new similarity measure, the "Hardy–Weinberg kernel", specifically conceived for incorporating prior knowledge during the study of genetic datasets of marker genotypes. The characteristic of "Hardy–Weinberg kernel" is that the similarity between genetic profiles is weighted by the estimates of gene frequencies at Hardy-Weinberg equilibrium in the population.

Methods

In order to compare the effectiveness of our similarity measure with respect to other "well established" kernels (Linear, Polynomial and Gaussian kernel), we applied SVM classification algorithms to two datasets (see below). For each classification problem SVM parameters were optimized through a cross validation procedure, while relevant features were selected via a backward – stepwise algorithm.

The Support Vector Machine classifier

SVMs represent a set of supervised learning methods that are used for classification problems. Viewing input data as two sets of vectors in an *N*-dimensional space, an SVM will construct a optimally separating hyperplane in that space, maximizing the margin between the two data sets. In order to define the maximal-margin hyperplane it is needed to solve a quadratic programming problem. In the next figure are shown two cases in which the two classes are linearly (Fig. 1a) and non linearly (Fig. 1b) separated. In many cases, it is more likely that two groups of objects are better separated by a non-linear hyperplane. To construct such a non-linear hyperplane, the data are mapped into a feature space by a kernel function. In this feature space a linear hyperplane is constructed as explained above, and then the data are mapped back into the original data space in which the separating hyperplane is not linear anymore (Fig. 2).

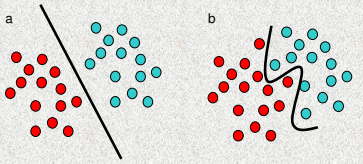


Figure 1. A Support Vector Machine performs classification by constructing an *N*-dimensional hyperplane that optimally separates the data into two categories. Linearly (a) and non linearly (b) separable cases are represented.

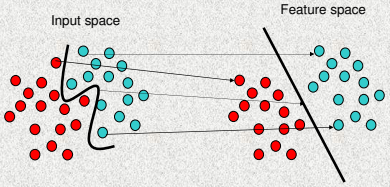


Figure 2. The objects in the input space are mapped into a feature space by an hypothetic kernel function. In this new setting the mapped objects are linearly separable.

Among the different kernel functions, the most commonly used are represented by the Polynomial, the Gaussian and the Sigmoid kernel functions.

The Hardy-Weinberg kernel

SVM classifiers cannot handle categorical data as our Single Nucleotide Polymorphisms (SNPs) data directly. A solution to this problem is to construct three binary dummy variables for each SNP, since SVMs can be applied to binary data. The first dummy variable of each SNP is set to 1 if both major alleles are present in the genotype, and to 0 otherwise. The second dummy variable of each SNP is set to 1 if both alleles are present in the genotype (heterozygous subjects), and to 0 otherwise. The third dummy variable of each SNP is set to 1 if both minor alleles are present in the genotype, and to 0 otherwise. At the end of the process, we obtain the "converted dataset". In order to explain this process, an example of genetic dataset and the relevant codifies are reported in Table 1 and Table 2, respectively.

Table 1. Example of a genetic dataset.

ID	SNP1	SNP2	SNP3
SAMPLE1	A1A1	B1B2	C2C2
SAMPLE2	A1A1	B1B2	C2C2
SAMPLE3	A1A2	B2B2	C1C1
SAMPLE4	A1A1	B1B1	C2C1

Table 2. The converted dataset obtained from the genetic data reported in Table 1.

ID	SNP1_A1A1	SNP1_A1A2	SNP1_A2A2	SNP2_B1B1	SNP2_B1B2	SNP2_B2B2	SNP3_C1C1	SNP3_C1C2	SNP3_C2C2
SAMPLE1	1	0	0	0	1	0	0	0	1
SAMPLE2	1	0	0	0	1	0	0	0	1
SAMPLE3	0	1	0	0	0	1	1	0	0
SAMPLE4	1	0	0	1	0	0	0	1	0

In such a way now we are able to handle genetic data (SNPs) into a SVM classifier. In order to introduce the Hardy-Weinberg kernel we suppose to compute a similarity measure between two subjects with respect to the genetic profiles reported in table 2 using a linear kernel function:

$$< SAMPLE1, SAMPLE4 >= (1,0,0,0,1,0,0,0,1)' \cdot (1,0,0,1,0,0,0,1,0) = 1$$

In this case the kernel function computes, for the entire set of the SNPs analyzed, the number of common genotypes between the pair of the subjects analyzed (dot product).

The observation that similarity between two subjects carrying a rare genotype have to be more important than the similarity between two subjects carrying a more frequent genotype suggest us to formulate the Hardy-Weinberg kernel. This new kernel computes the similarity measure between the same pair of subjects in a two steps procedure. First, for each column *i* of the converted dataset a scale factor *fi* is calculated. The scale factor fi corresponds to the genotypic frequency F of the genotype represented by the column i, normalized for the maximum genotypic frequency of the respective polymorphism. In this context "genotypic frequency" always indicates the genotypic frequency in the population. For example, for a generic SNP with three genotypes, namely *A1A1*, *A1A2* and A2A2, the scale factors for the respective columns in the converted dataset will be:

$$f_{A1A1} = \frac{F_{A1A1}}{\max(F_{A1A1}, F_{A1A2}, F_{A2A2})} \quad f_{A1A2} = \frac{F_{A1A2}}{\max(F_{A1A1}, F_{A1A2}, F_{A2A2})} \quad f_{A2A2} = \frac{F_{A2A2}}{\max(F_{A1A1}, F_{A1A2}, F_{A2A2})}$$

Once terminated the first step, all the values in the converted dataset are subdivided by the scale factor of the respective column. With the Hardy–Weinberg kernel, the previous example becomes:

$$\langle SAMPLE1 \cdot SAMPLE2 \rangle =$$

$$\langle (1/f_{A1A1}, 0,0,0,1/f_{A1A2}, 0,0,0,1/f_{C1C2}) \cdot (1/f_{A1A1}, 0,0,1/f_{A1A1}, 0,0,0,1/f_{C1C2}, 0) \rangle = \left(\frac{1}{f_{A1A1}}\right)^2 = \left(\frac{1}{\frac{F_{A1A1}}{\max(F_{A1A1}, F_{A1A2}, F_{A2A2})}}\right)^2$$

In other words, the Hardy-Weinberg kernel is a similarity measure that, in the frame of SVM classifiers, is able to compare genotypic profiles by weighting the similarities between different subjects for the inverse of the genotypic frequencies at the Hardy-Weinberg equilibrium

Datasets

Dataset 1 had been collected in order to investigate the influence of the variability of 70 independent SNPs on survival at old age in Italian population. The products of the genes analyzed are involved in the pathways of insulin, IGF-1, in the inflammatory response and in the oxidative stress. 47 long lived individuals (19 males and 28 females, median age 98 years) recruited from Calabrian population (Southern Italy) have been genotyped by using an APEX (Arrayed Primer EXtension) technology. Control group has been obtained by using a simulation approach. One hundred control samples have been generated by using the genotypic frequencies of Caucasian population obtained from the dbSNP database available at web site http://www.ncbi.nlm.nih.gov/sites/entrez. Finally, assembling cases and control groups, 100 datasets have been generated. Then, each dataset has been analyzed by applying the two different kernel functions into the SVM classifier.

Dataset 2 had been collected in order to investigate the role of about 200 SNPs on the risk to develop colorectal cancer. A sample of 377 cases and 329 sex and age matched controls has been recruited from the Spanish population. The products of the genes analyzed are involved in the metabolism of dietary carcinogens and xenobiotics, in the repair and replication of DNA and in the apoptotic process. Also in this case, genotypic analyses have been carried out by using an APEX approach.

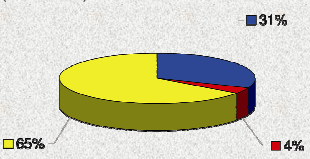
Genotypic analyses

Arrayed Primer EXtension consists of a sequencing reaction primed by an oligonucleotide anchored with its 5' end to a glass slide, and terminating just one nucleotide before the polymorphic site. A DNA polymerase extends the oligonucleotide by adding one fluorescently labelled dideoxynucleotide triphosphate complementary to the variant base. Reading the incorporated fluorescence identifies the base in the target sequence. PCR products were purified and concentrated using Millipore Y30 columns. In order to allow better hybridisation with the arrayed oligonucleotides, the PCR products were reduced in size by fragmentation. The mixture was placed quickly onto the spotted slide and incubated. Slides were imaged by a Genorama-003 four-colour detector (Asper Biotech, Tartu, Estonia). Four images were analysed, each corresponding to a fluorochrome (i.e. a base). Fluorescence intensities at each position was measured and converted to base calls according to the Genorama image analysis and genotyping software (Asper Biotech, Tartu, Estonia).

Results

We inserted the two kernel functions, the linear and the Hardy-Weinberg kernels, into the SVM classifier and, successively, we evaluated the relevant performances in terms of AUC values. SVM parameters were optimized through a cross validation procedure.

Dataset 1. The next figure shows the percentages of simulated datasets in which the performance of Hardy-Weinberg kernel matches (yellow), overcomes (blue) or shows a poor performance (red) than those obtained using a linear kernel into a SVM classifier on the data of longevity. In total 100 datasets have been simulated (see Methods).



- Hardy-Weinberg kernel shows higher performances than those of the linear kernel in 31% of the simulated datasets.
- Hardy-Weinberg kernel matches the performances obtained using a linear kernel in 65% of the simulated datasets.
- Hardy-Weinberg kernel shows lower performances than those of the linear kernel in 4% of the simulated datasets.

Dataset 2. The next table reports the AUC values obtained applying an SVM classifier on the data of colorectal cancer using the linear and the Hardy-Weinberg kernels.

Table 3. AUC values obtained applying an SVM classifier on the data of colorectal cancer using the linear and the Hardy-Weinberg kernels.

Kernel function	Performance (AUC)	P-value*
Hardy-Weinberg Kernel	64.4%	<0.001
Linear Kernel	62.6%	

*P-value refers to the comparison between ten AUC values obtained from a ten-fold cross-validation procedure

Conclusions

We set up a new kernel, the Hardy-Weinberg kernel, specifically able to handle genetic data into SVM. In most cases the Hardy-Weinberg kernel performances match or overcome the performances of the linear kernel. Other datasets (both with nuclear polymorphisms and mtDNA sequences) are currently under study to verify the goodness of this new similarity measure. Hardy-Weinberg kernel may represent a valuable tool for the case-control studies carried out with high throughput genotyping.