Chapter 14
GA-Based Data Mining Applied to Genetic Data for the Diagnosis of Complex Diseases

Vanessa Aguiar
University of A Coruña, Spain

Jose A. Seoane
University of A Coruña, Spain

Ana Freire
University of A Coruña, Spain

Ling Guo
University of A Coruña, Spain

ABSTRACT

A new algorithm is presented for finding genotype-phenotype association rules from data related to complex diseases. The algorithm was based on genetic algorithms, a technique of evolutionary computation. The algorithm was compared to several traditional data mining techniques and it was proved that it obtained better classification scores and found more rules from the data generated artificially. It also obtained similar results when using some UCI Machine Learning datasets. In this chapter it is assumed that several groups of Single Nucleotide Polymorphisms (SNPs) have an impact on the predisposition to develop a complex disease like schizophrenia. It is expected to validate this in a short period of time on real data.

INTRODUCTION

Complex diseases are those that result from the interaction of multiple factors, usually including both genetic and environmental factors (Risch, 2000). Due to their nature, it is hard to establish a relationship between a gene and the disease. In general, this type of disease is caused by combination of effects of several sets of Single Nucleotide Polymorphisms (SNPs) which, separately, have a low effect. There is a high prevalence and impact of complex diseases like cancer, mental disorders and cardiovascular diseases. This situation has a

DOI: 10.4018/978-1-61520-893-7.ch014

high repercussion on the costs of hospitals and, therefore, on the costs of the national health system.

A SNP (Den Dunnen & Antonarakis, 2000) is a single nucleotide site where two (of four) different nucleotides occur in a high percentage of the population, that is, at least in 1% of the population. Since there exist 14 million of SNPs in human beings then a huge amount of data obtained from DNA genotyping needs to be dealt with, thus many variables have to be taken into account.

This data can be analysed carrying out association studies. In a genetic association study, the frequency of a SNP variant in people affected by the same disease is compared to the frequency of a SNP variant in healthy people (control population). There has to be no familiar relationship between these subjects, they have to belong to the same ethnic group and have the same geographic origin.

Carrying out such studies is expensive, mostly due to the cost of genotyping. Genotyping is the process of determining the genotype of an individual using a biological test. In Spain, for example, the cost of genotyping 74 SNPs for 720 samples reaches nearly 8.000€. The accuracy rate of the technologies used for this purpose ranges between 85-98%, depending on which one has been chosen. The technology used is chosen depending on the approach and purpose of the study and the number of SNPs to be genotyped. Not having an accuracy rate of 100% will make the analysis of genetic data more difficult as there will be missing data.

An important challenge that molecular association study faces in the post genomic era is to understand the inter-connections between networks of genes and their products. These networks are initiated and regulated by a variety of environmental changes. The variety of genotype definitions leads to an increase of the number of tests that need to be run and also involves a large amount of comparisons. Non-reproducibility of many results obtained in several studies has led to criticism of association studies.

SNP data and haplotypes used in association studies of complex diseases have three main characteristics which represent important challenges in data analysis. These characteristics are: complexity, heterogeneity and a constantly evolving nature. In addition to this, this type of data is large, redundant, diverse and distributed.

It is heterogeneous in the sense that it involves a large amount of data types, including categorical and continuous data, sequences, as well as temporal data, incomplete and missing data. There is a lot of redundancy in SNP and haplotype databases. This type of data is very dynamic and evolves continuously. Not only the data but also the schema evolves, which means that it requires special knowledge when designing modelling techniques. Finally, SNP and haplotype data is complex and has intrinsic features and subtle patterns, in the sense that it is very rich in associated complex phenotype traits or common multifactor diseases.

In complex diseases, in general, the combination of certain genes predisposes to develop a disease and the environmental factors are those which increase the impact of these genes in the disease development. This is known as epistasis or epistatic effect. In addition, environmental factors, which at the population level seem to have only a moderate impact, might have higher risks in subpopulations with certain genetic predispositions. There are major methodological challenges in the study of gene-gene and gene-environment interactions. Another important challenge is to study large datasets in order to identify combinations of SNPs which interact increasing the predisposition to develop a certain complex disease. Thus, there is a need to develop methods capable of performing a massive analysis of SNP data related to complex diseases beyond that of traditional statistical approaches.

Hence, the objective of this chapter is to develop an algorithm that will analyse data obtained from genotyping as part of an association study.