Hypothesis Testing with Classifier Systems for Rule-Based Risk Prediction

Flavio Baronti and Antonina Starita

Dipartimento di Informatica, Università di Pisa
Largo B. Pontecorvo, 3—56127 Pisa, Italy
{baronti, starita}@di.unipi.it

Abstract. Analysis of medical datasets has some specific requirements not always fulfilled by standard Machine Learning methods. In particular, heterogeneous and missing data must be tolerated, the results should be easily interpretable. Moreover, with genetic data, often the combination of two or more attributes leads to non-linear effects not detectable for each attribute on its own. We present a new ML algorithm, HCS, taking inspiration from learning classifier systems, decision trees and statistical hypothesis testing. We show the results of applying this algorithm to a well-known benchmark dataset, and to HNSCC, a dataset studying the connection between smoke and genetic patterns to the development of oral cancer.

1 Introduction

Medical research is shifting its focus from populations to individuals. It was already evident that people react differently to the same stimuli (diseases, therapies). The discovery of DNA and the advent of genetic profiling suggested it was possible to track down these differences to their root causes. The genetic profile of a person should have pinpointed her exact reaction to diseases and medicines; unfortunately, this objective is still very far in the future. We can describe two main aspects of genetic understanding which make the goal difficult to reach. The first one is *genome size*. As everyday experience clearly shows, every person is unique. From the genetic point of view, this can be explained through the huge number of different genes in the human genome: it is very unlikely that two people have exactly the same allelic variants of each gene. In principle, this makes generalization impossible: we will never find "the same situation". This problem can be solved however by understanding the effect each gene has, and by considering only those genes involved in the disease being studied. If for a particular disease we can reduce the number of involved genes to a manageable number, finding the same situation becomes possible. The second difficulty is *gene interaction*. Traditional medical analyses typically assume independence of the causing factors, and linearity of their combined effect. These assumptions appear wrong in genetic research: rarely a single gene variant has a direct effect on the outcome; genes work together, and often only their combined effect shows

E. Marchiori, J.H. Moore, and J.C. Rajapakse (Eds.): EvoBIO 2007, LNCS 4447, pp. 24–34, 2007.

a significant impact on the outcome. This observation requires different analysis techniques to be used, able to deal with non-linearity and factors dependence.

In this paper we will present a new algorithm, developed in order to analyze data collected to study the effects of genetic variants on development of Head and Neck Squamous Cell Carcinoma (HNSCC), a kind oral cancer very common among smokers. Smoking is clearly a very important risk factor; observation however shows that there exist more sensitive people, which develop cancer with little to no smoking, and more resistant people, which do not develop cancer although smoking a lot. This difference could be explained by different treatment of carcinogens by the organism, regulated by different genetic variants. The analysis was performed through a new machine learning (ML) algorithm, loosely based upon learning classifier systems (LCS) research [1], called HCS (Hypothesis testing with Classifier Systems). Its main aim is to identify subsets of the whole dataset where the risk factor is significantly different from the global risk. Differently from most ML algorithms, HCS loosens the requirement of accuracy of its prediction, allowing to output a risk value instead of an exact classification. The main focus is instead generality of the prediction: we will show how this shift of goals was beneficial to the importance of results.

2 Problem Description

The data set we analyzed (originally presented in [2]) was designed to explore the influence of genotype on the chance to develop head and neck squamous cell carcinoma (HNSCC). It is already well-known that this kind of cancer is associated with smoking and alcohol-drinking habits, it is more common among males and its incidence increases with age. The individual risk however could be modified by genetic factors; therefore genotype information, regarding eleven genes involved with carcinogen-metabolizing (`CCND1`, `NQO1`, `EPHX1`, `CYP2A6`, `CYP2D6`, `CYP2E1`, `NAT1`, `NAT2`, `GSTP1`) and DNA repair systems (`OGG1`, `XPD`) was provided by molecular testing.

Nine of these genes have two allelic variants; let's call them a_1 and a_2. Since the DNA contains two copies of each gene, there exist three possible combinations: a_1a_1, a_2a_2 (the homozygotes) and a_1a_2 (the heterozygote — order does not matter). The homozygotes where represented with values 0 and 2, while the heterozygote with 1. Due to dominance, for the examined genes the heterozygote is equivalent to one of the homozygotes; however, for many of the considered genes this dominant effect is not known. So class 1 is either equivalent to class 0, or to class 2. The remaining two genes (`NAT1` and `NAT2`) have 4 allelic variants, which result in 9 combinations; they were sorted by their activity level, and put on an integer scale from 0 to 8.

The full data consists of 355 records, with 124 positive elements (HNSCC patients) and 231 negative (controls). Each record reports the person's gender, age, total smoke and alcohol consumption, gene values, and a boolean target value which specifies whether he had cancer when the database was compiled or not. The data was collected in different periods between 1997 and 2003; this