Tuning ReliefF for Genome-Wide Genetic Analysis

Jason H. Moore and Bill C. White

Dartmouth College, One Medical Center Drive, Lebanon, NH 03756
Jason.H.Moore@dartmouth.edu, Bill.C.White@dartmouth.edu

Abstract. An important goal of human genetics is the identification of DNA sequence variations that are predictive of who is at risk for various common diseases. The focus of the present study is on the challenge of detecting and characterizing nonlinear attribute interactions or dependencies in the context of a genome-wide genetic study. The first question we address is whether the ReliefF algorithm is suitable for attribute selection in this domain. The second question we address is whether we can improve ReliefF for selecting important genetic attributes. Using simulated genetic datasets, we show that ReliefF is significantly better than a naïve chi-square test of independence for selecting two interacting attributes out of 10^3 candidates. In addition, we show that ReliefF can be improved in this domain by systematically removing the worst attributes and re-estimating ReliefF weights. Our simulation studies demonstrate that this new Tuned ReliefF (TuRF) algorithm is significantly better than ReliefF.

1 Introduction

1.1 The Problem Domain: Human Genetics

Biological and biomedical sciences are undergoing an information explosion and an understanding implosion. That is, our ability to generate data is far outpacing our ability to interpret it. This is especially true in the domain of human genetics where it is now technically feasible to measure thousands of DNA sequence variations from across the human genome. For the purposes of this paper we will focus exclusively on the single nucleotide polymorphism or SNP which is a single nucleotide or point in the DNA sequence that differs among people. It is anticipated that at least one SNP occurs approximately every 100 nucleotides across the $3 * 10^9$ nucleotide human genome. An important goal in human genetics is to determine which of the many thousands of SNPs are useful for predicting who is at risk for common diseases such as prostate cancer, cardiovascular disease, or bipolar depression. This "genome-wide" approach is expected to revolutionize the genetic analysis of common human diseases [1], [2].

The charge for computer science and bioinformatics is to develop algorithms for the detection and characterization of those SNPs that are predictive of human health and disease. Success in this genome-wide endeavor will be difficult

E. Marchiori, J.H. Moore, and J.C. Rajapakse (Eds.): EvoBIO 2007, LNCS 4447, pp. 166–175, 2007.

due to nonlinearity in the genotype-to-phenotype mapping relationship that is due, in part, to epistasis or nonadditive gene-gene interactions. Epistasis was recognized by Bateson [3] nearly 100 years ago as playing an important role in the mapping between genotype and phenotype. Today, this idea prevails and epistasis is believed to be a ubiquitous component of the genetic architecture of common human diseases [4]. As a result, the identification of genes with genotypes that confer an increased susceptibility to a common disease will require a research strategy that embraces, rather than ignores, this complexity [4], [5], [6]. The implication of epistasis from a data mining point of view is that SNPs need to be considered jointly in learning algorithms rather than individually. Because the mapping between the attributes and class is nonlinear, the concept difficulty is high. The challenge of modeling attribute interactions has been previously described [7]. Due to the combinatorial magnitude of this problem, intelligent feature selection strategies are needed. The goal of this paper is to evaluate the ReliefF algorithm as a statistical filter in this domain.

1.2 A Simple Example of the Concept Difficulty

Epistasis can be defined as biological or statistical [5]. Biological epistasis occurs at the cellular level when two or more biomolecules physically interact. In contrast, statistical epistasis occurs at the population level and is characterized by deviation from additivity in a linear mathematical model. Consider the following simple example of statistical epistasis in the form of a penetrance function. Penetrance is simply the probability (P) of disease (D) given a particular combination of genotypes (G) that was inherited (i.e. $P[D|G]$). A single genotype is determined by one allele (i.e. a specific DNA sequence state) inherited from the mother and one allele inherited from the father. For most single nucleotide polymorphisms or SNPs, only two alleles (encoded by A or a) exist in the biological population. Therefore, because the order of the alleles is unimportant, a genotype can have one of three values: AA, Aa or aa. The model illustrated in Table 1 is an extreme example of epistasis. Let's assume that genotypes AA, aa, BB, and bb have population frequencies of 0.25 while genotypes Aa and Bb have frequencies of 0.5 (values in parentheses in Table 1). What makes this model interesting is that disease risk is dependent on the particular combination of genotypes inherited. Individuals have a very high risk of disease if they inherit Aa or Bb but not both (i.e. the exclusive OR function). The penetrance for each individual genotype in this model is 0.5 and is computed by summing the products of the genotype frequencies and penetrance values. Thus, in this model there is no difference in disease risk for each single genotype as specified by the single-genotype penetrance values. This model was first described by Li and Reich [8]. Heritability or the size of the genetic effect is a function of these penetrance values. In this model, the heritability is maximal at 1.0 because the probability of disease is completely determined by the genotypes at these two DNA sequence variations. This is a special case where all of the heritability is due to epistasis. As Freitas reviews [7], this general class of problems has high concept difficulty.