Chapter XXXVIII
Multitarget Classifiers for Mining in Bioinformatics

Diego Liberati
Istituto di Elettronica e Ingegneria dell'Informazione e delle Telecomunicazioni Consiglio Nazionale delle Ricerche Politecnico di Milano, Italy

ABSTRACT

Building effective multitarget classifiers is still an on-going research issue: this chapter proposes the use of the knowledge gleaned from a human expert as a practical way for decomposing and extend the proposed binary strategy. The core is a greedy feature selection approach that can be used in conjunction with different classification algorithms, leading to a feature selection process working independently from any classifier that could then be used. The procedure takes advantage from the Minimum Description Length principle for selecting features and promoting accuracy of multitarget classifiers. Its effectiveness is asserted by experiments, with different state-of-the-art classification algorithms such as Bayesian and Support Vector Machine classifiers, over dataset publicly available on the Web: gene expression data from DNA micro-arrays are selected as a paradigmatic example, containing a lot of redundant features due to the large number of monitored genes and the small cardinality of samples. Therefore, in analysing these data, like in text mining, a major challenge is the definition of a feature selection procedure that highlights the most relevant genes in order to improve automatic diagnostic classification.

INTRODUCTION

As stressed by recent literature (Mukherjee, 2003; Statnikov et al., 2005) many classifiers have their own limitations when used for multitarget classification, i.e. predicting more than two different classes. The prior knowledge gleaned from a human expert is employed as a support to overcome the problems faced by multitarget classification: a divide-and-conquer approach allows to decompose the original multitarget classification problem into a set of binary classification problems. Such decomposition is

performed by using a decision tree that is based on previous knowledge, experience and observation (Yeoh et al., 2002) and provides an inference schema reflecting the human decision making process.

This is of paramount importance when dealing with patho-physiological problems, like the process of selecting the most important genes, i.e. the minimal set of genes allowing to build efficient classifiers, from micro-array data (Golub et al., 1999; Guyon et al., 2002; Blum and Lanley, 1997), as well as when selecting the most interesting features in text mining.

In order to improve the performance of learning algorithms (Kahn et al., 2001; Golub et al., 1999) and avoid over-fitting, it is of paramount importance to reduce the dimensionality of the data by deleting unsuitable attributes (Witten and Frank, 2005).

BACKGROUND

The common practice in such increasingly important bioinformatics field is to employ a range of accessible methodologies that can be broadly classified into three categories:

- Classification methods based on global gene expression analysis (Golub et al., 1999; Alizadeh et al., 2000; Ross et al., 2000) specifically aimed at applying a single technique to a specific gene expression dataset;
- Traditional statistical approaches such as Principal Component Analysis (Liberati et al., 2005; Garatti et al., 2007), discriminant analysis (Nguyen and Rocke, 2002) or Bayesian decision theory (Bosin et al., 2006);
- Machine learning techniques such as neural (Tung and Quek, 2005; Khan et al., 2001) and logical (Muselli and Liberati, 2002) networks, decision trees and Support Vector Machines (SVM) (Guyon et al., 2002; Furey et al., 2001; Valentini, 2002).

Nonetheless, (Statnikov et al., 2005) reported that such results lack a consistent and systematic approach as they validate their methods differently, on different public datasets and on different limited sets of features. Dudoit (2002) and colleagues have compared the performance of various micro-array data classification methods, and a recent extensive comparison (Lee et al., 2005) provides some additional insights. The relevance of good feature selection methods has been discussed by Guyon (2002) and colleagues with special emphasis on over-fitting, but the recommendations in literature do not give evidence for a single best method for either the classification of micro-array data, or at least their feature selection. It has also been pointed out (Tung and Quek, 2005) that often classifiers work as black boxes, the decision making process being not intuitive to the human cognitive process and, more importantly, the knowledge extracted by these classifiers from the numerical training not being easy to be understood and then assessed.

In order to overcome such drawbacks, an approach more driven by data in feature selection, not neglecting the available domain knowledge, makes it possible to emulate the human style of reasoning and decision making when solving complex problems.

The adoption of the Minimum Description Length (MDL) principle (Barron et al., 1998) is proposed for both selecting features and comparing classifiers. Any regularity in the data can in fact be used to compress the data themselves. Being data compression equivalent to a kind of probabilistic prediction, MDL methods can be interpreted as searching for a model with good predictive performance on unseen