A Genetic Embedded Approach for Gene Selection and Classification of Microarray Data

Jose Crispin Hernandez Hernandez, Béatrice Duval, and Jin-Kao Hao

LERIA, Université d'Angers,
2 Boulevard Lavoisier, 49045 Angers, France
{josehh,bd,hao}@info.univ-angers.fr

Abstract. Classification of microarray data requires the selection of subsets of relevant genes in order to achieve good classification performance. This article presents a genetic embedded approach that performs the selection task for a SVM classifier. The main feature of the proposed approach concerns the highly specialized crossover and mutation operators that take into account gene ranking information provided by the SVM classifier. The effectiveness of our approach is assessed using three well-known benchmark data sets from the literature, showing highly competitive results.

Keywords: Microarray gene expression, Feature selection, Genetic Algorithms, Support vector machines.

1 Introduction

Recent advances in DNA microarray technologies enable to consider molecular cancer diagnosis based on gene expression. Classification of tissue samples from gene expression levels aims to distinguish between normal and tumor samples, or to recognize particular kinds of tumors [9,2]. Gene expression levels are obtained by cDNA microarrays and high density oligonucleotide chips, that allow to monitor and measure simultaneously gene expressions for thousands of genes in a sample. So, data that are currently available in this field concern a very large number of variables (thousands of gene expressions) relative to a small number of observations (typically under one hundred samples). This characteristic, known as the "curse of dimensionality", is a difficult problem for classification methods and requires special techniques to reduce the data dimensionality in order to obtain reliable predictive results.

Feature selection aims at selecting a (small) subset of informative features from the initial data in order to obtain high classification accuracy [11]. In the literature there are two main approaches to solve this problem: the filter approach and the wrapper approach [11]. In the filter approach, feature selection is performed without taking into account the classification algorithm that will be applied to the selected features. So a filter algorithm generally relies on a relevance measure that evaluates the importance of each feature for the classification task. A feasible approach to filter selection is to rank all the features

E. Marchiori, J.H. Moore, and J.C. Rajapakse (Eds.): EvoBIO 2007, LNCS 4447, pp. 90–101, 2007.

according to their interestingness for the classification problem and to select the top ranked features. The feature score can be obtained independently for each feature, as it is done in [9] which relies on correlation coefficients between the class and each feature. The drawback of such a method is to score each feature independently while ignoring the relations between the features.

In contrast, the wrapper approach selects a subset of features that is "optimized" by a given classification algorithm, *e.g.* a SVM classifier [5]. The classification algorithm, that is considered as a black box, is run many times on different candidate subsets, and each time, the quality of the candidate subset is evaluated by the performance of the classification algorithm trained on this subset. The wrapper approach conducts thus a search in the space of candidate subsets. For this search problem, genetic algorithms have been used in a number of studies [15,14,6,4].

More recently, the literature also introduced embedded methods for feature selection. Similar to wrapper methods, embedded methods carry out feature selection as a part of the training process, so the learning algorithm is no more a simple black box. One example of an embedded method is proposed in [10] with recursive feature elimination using SVM (SVM-RFE).

In this paper, we present a novel embedded approach for gene selection and classification which is composed of two main phases. For a given data set, we carry out first a pre-selection of genes based on filtering criteria, leading to a reduced gene subset space. This reduced space is then searched to identify even smaller subsets of predictive genes which are able to classify with high accuracy new samples. This search task is ensured by a specialized Genetic Algorithm which uses (among other things) a SVM classifier to evaluate the fitness of the candidate gene subsets and problem specific genetic operators. Using SVM to evaluate the fitness of the individuals (gene subsets) is not a new idea. Our main contribution consists in the design of semantically meaningful crossover and mutation operators which are fully based on useful ranking information provided by the SVM classifier. As we show in the experimentation section, this approach allows us to obtain highly competitive results on three well-known data sets.

In the next Section, we recall three existing filtering criteria that are used in our pre-selection phase and SVM that is used in our GA. In Section 3, we describe our specialized GA for gene selection and classification. Experimental results and comparisons are presented in Section 4 before conclusions are given in Section 5.

2 Basic Concepts

2.1 Filtering Criteria for Pre-selection

As explained above, microarray data generally concern several thousands of gene expressions. It is thus necessary to pre-select a smaller number of genes before applying other search methods. This pre-selection can be performed by using simply a classical filter method that we recall in this section. The following