Information Sciences 282 (2014) 111–135

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

A review of microarray datasets and applied feature selection methods

V. Bolón-Canedo [a,*], N. Sánchez-Maroño [a], A. Alonso-Betanzos [a], J.M. Benítez [b], F. Herrera [b,c]

[a] Department of Computer Science, Universidade de A Coruña, 15071 A Coruña, Spain
[b] Department of Computer Science and Artificial Intelligence, Universidad de Granada, 18071 Granada, Spain
[c] Faculty of Computing and Information Technology – North Jeddah, King Abdulaziz University, 21589 Jeddah, Saudi Arabia

ARTICLE INFO

Article history:
Received 13 November 2013
Received in revised form 4 March 2014
Accepted 20 May 2014
Available online 14 June 2014

Keywords:
Feature selection
Microarray data
Unbalanced data
Dataset shift

ABSTRACT

Microarray data classification is a difficult challenge for machine learning researchers due to its high number of features and the small sample sizes. Feature selection has been soon considered a *de facto* standard in this field since its introduction, and a huge number of feature selection methods were utilized trying to reduce the input dimensionality while improving the classification performance. This paper is devoted to reviewing the most up-to-date feature selection methods developed in this field and the microarray databases most frequently used in the literature. We also make the interested reader aware of the problematic of data characteristics in this domain, such as the imbalance of the data, their complexity, or the so-called dataset shift. Finally, an experimental evaluation on the most representative datasets using well-known feature selection methods is presented, bearing in mind that the aim is not to provide the best feature selection method, but to facilitate their comparative study by the research community.

1. Introduction

During the last two decades, the advent of DNA microarray datasets has stimulated a new line of research both in bioinformatics and in machine learning. This type of data is used to collect information from tissue and cell samples regarding gene expression differences that could be useful for disease diagnosis or for distinguishing specific types of tumor. Although there are usually very small samples (often less than 100 patients) for training and testing, the number of features in the raw data ranges from 6000 to 60,000, since it measures the gene expression en masse. A typical classification task is to separate healthy patients from cancer patients based on their gene expression "profile" (binary approach). There are also datasets in which the goal is to distinguish among different types of tumors (multiclass approach), making the task even more complicated.

Therefore, microarray data pose a serious challenge for machine learning researchers. Having so many fields relative to so few samples creates a high likelihood of finding "false positives" due to chance (both in finding relevant genes and in building predictive models) [94]. It becomes necessary to find robust methods to validate the models and assess their likelihood. Furthermore, additional experimental complications (like noise and variability) render the analysis of microarray data an exciting domain [98].

Several studies have shown that most genes measured in a DNA microarray experiment are not relevant in the accurate classification of different classes of the problem [46]. To avoid the problem of the "curse of dimensionality" [62], feature

* Corresponding author. Tel.: +34 981 167000.
E-mail addresses: vbolon@udc.es (V. Bolón-Canedo), nsanchez@udc.es (N. Sánchez-Maroño), ciamparo@udc.es (A. Alonso-Betanzos), j.m.benitez@decsai.ugr.es (J.M. Benítez), herrera@decsai.ugr.es (F. Herrera).

http://dx.doi.org/10.1016/j.ins.2014.05.042
0020-0255/

(gene) selection plays a crucial role in DNA microarray analysis, which is defined as the process of identifying and removing irrelevant features from the training data, so that the learning algorithm focuses only on those aspects of the training data useful for analysis and future prediction [50]. There are usually three varieties of feature selection methods: filters, wrappers and embedded methods. While wrapper models involve optimizing a predictor as part of the selection process, filter models rely on the general characteristics of the training data to select features independent of any predictor. The embedded methods generally use machine learning models for classification, and then an optimal subset of features is built by the classifier algorithm. Of course, the interaction with the classifier required by wrapper and embedded methods comes with an important computational burden (more important in the case of wrappers). In addition to this classification, feature selection methods may also be divided into univariate and multivariate types. Univariate methods consider each feature independently of other features, a drawback that can be overcome by multivariate techniques that incorporate feature dependencies to some degree, at the cost of demanding more computational resources [26].

Feature selection as a preprocessing step to tackle microarray data has rapidly become indispensable among researchers, not only to remove redundant and irrelevant features, but also to help biologists identify the underlying mechanism that relates gene expression to diseases. This research area has received significant attention in recent years (most of the work has been published in the last decade), and new algorithms have emerged as alternatives to the existing ones. However, when a new method is proposed, there is a lack of standard state-of-the-art results to perform a fair comparative study. Furthermore, there is a broad suite of microarray datasets to be used in the experiments, some of which even have the same name, but the number of samples or characteristics are different in different studies, which makes this task more complicated.

The main goal of the research presented here is to provide a review of the existing feature selection methods developed to be applied to DNA microarray data. In addition to this, we pay attention to the datasets used, their intrinsic data characteristics and the behavior of classical feature selection algorithms available in data mining software tools used for microarray data. In this manner, the reader can be aware of the particularities of this type of data as well as its problematics, such as the imbalance of the data, their complexity, the presence of overlapping and outliers, or the so-called dataset shift. These problematics render the analysis of microarray data an interesting domain.

We have designed an experimental study in such a way that we can draw well-founded conclusions. We use a set of nine two-class microarray datasets, which suffer from problems such as class imbalance, overlapping or dataset shift. Some of these datasets were originally divided into training and test datasets, so a holdout validation is performed on them. For the remaining datasets, we choose to evaluate them with a k-fold cross-validation, since it is a common choice in the literature [81,107,86,101,31,105,125]. However, it has been shown that cross-validation can potentially introduce dataset shift, so we include another strategy to create the partitioning, called *Distribution optimally balanced stratified cross-validation* (DOB-SCV) [84]. We consider C4.5, Support Vector Machine (SVM) and naive Bayes as classifiers, and we use classification accuracy, sensitivity and specificity on the test partitions as the evaluation criteria.

The remainder of the paper is organized as follows: Section 2 introduces the background and the first attempts to deal with microarray datasets. In Section 3 we review the state of the art on feature selection methods applied to this type of data, including the classical techniques (filters, embedded and wrappers) as well as other more recent approaches. Next, Section 4 is focused on the particularities of the datasets, from providing a summary of the characteristics of the most famous datasets used in the literature and existing repositories to the analysis of the inherent problematics of microarray data, such as the small-sample size, the imbalance of the data, the dataset shift or the presence of outliers. In Section 5 we present an experimental study of the most significant algorithms and evaluation techniques. A deep analysis of the findings of this study is also provided. Finally, in Section 6, we make our concluding remarks.

2. Background: the problem and first attempts

All cells have a nucleus, and inside this nucleus there is DNA, which encodes the "program" for future organisms. DNA has coding and non-coding segments. The coding segments, also known as genes, specify the structure of proteins, which do the essential work in every organism. Genes make proteins in two steps: DNA is transcribed into mRNA and then mRNA is translated into proteins. Advances in molecular genetics technologies, such as DNA microarrays, allow us to obtain a global view of the cell, with which it is possible to measure the simultaneous expression of tens of thousands of genes [94]. Fig. 1 displays the general process of acquiring the gene expression data from a DNA microarray. These gene expression profiles can be used as inputs to large-scale data analysis, for example, to increase our understanding of normal and diseased states.

Microarray datasets began to be dealt with at the end of the nineties. Soon feature (gene) selection was considered a *de facto* standard in this field. Further work was carried out at the beginning of the 2000s [98]. The univariate paradigm, which is fast and scalable but which ignores feature dependencies, has dominated the field during the 2000s [36,74,72]. However, there were also attempts to tackle microarray data with multivariate methods, which are able to model feature dependencies, but at the cost of being slower and less scalable than univariate techniques [26]. Apart from the application of multivariate filter methods [34,126,121,45], the microarray problem was also treated with more complex techniques such as wrappers and embedded methods [22,63,60,97].

So far we have briefly described the state-of-the-art of microarray data classification during its infancy. The next section is dedicated to reviewing the most up-to-date feature selection methods applied to this type of data.