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.

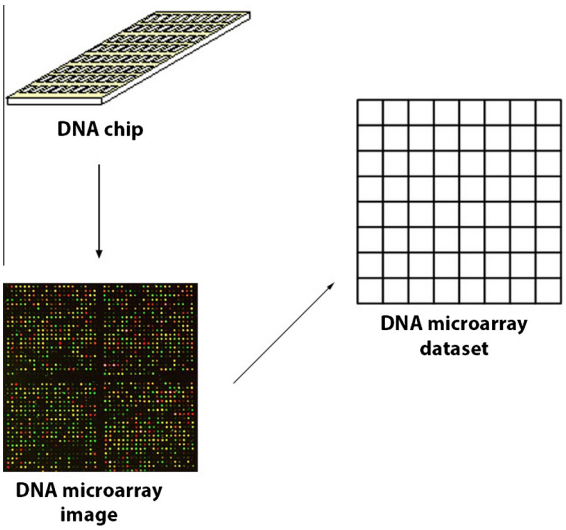

Fig. 1. General process of acquiring the gene expression data from DNA microarray.

3. Algorithms for feature selection on microarray data: a review

Feature selection methods are constantly emerging and, for this reason, there is a wide suite of methods that deal with microarray gene data. The aim of this section is to present those methods developed in the last few years. Traditionally, the most employed gene selection methods fall into the filter approach (see Section 3.1). Most of the novel filter methods proposed are based on information theory, although issues such as robustness or division in multiple binary problems are emerging topics. Discretization as a step prior to feature selection has also received some degree of attention. On the other hand, and due to the heavy computational consumption of resources and the high risk of overfitting, the wrapper approach has been largely avoided in the literature (see Section 3.2). Although the embedded model had not received sufficient attention during the infancy of microarray data classification, several proposals have emerged in recent years, as reported in Section 3.3. It is also worth noticing that the review of up-to-date literature has shown a tendency to mix algorithms, either in the form of hybrid methods or ensemble methods. Also, it is well-known that genes interact with each other through gene regulative networks, so clustering methods have also been proposed. These novel approaches will be described in Section 3.4.

The interested reader may find some works that review the feature selection methods used most in this field. Saeys et al. [98] provide a basic taxonomy of classical feature selection techniques and discuss their use in a number of bioinformatics applications. Lazar et al. [70] present a survey focused on filter feature selection methods in a unified framework, using standardized notations in order to reveal their technical details and to highlight their common characteristics as well as their particularities. Within the family of embedded methods, Ma and Huang [79] review some penalized feature selection methods, which are shown to be applicable to high-dimensional bioinformatics data.

3.1. Filters

Filter methods evaluate the goodness of gene subsets by observing only the intrinsic data characteristics (i.e. statistical measures), in which typically a single gene or a subset of genes is evaluated against the class label. Classical filter methods are usually applied to microarray data, such as Correlation Feature Selection (CFS), Fast Correlation-Based Filter (FCBF), ReliefF, or the consistency-based filter [98]. A brief description of these classical filters follows:

- **Correlation-based Feature Selection** (CFS) is a simple multivariate filter algorithm that ranks feature subsets according to a correlation based heuristic evaluation function [52]. The bias of the evaluation function is toward subsets that contain features that are highly correlated with the class and uncorrelated with each other. Irrelevant features should be ignored because they will have low correlation with the class. Redundant features should be screened out as they will be highly correlated with one or more of the remaining features. The acceptance of a feature will depend on the extent to which it predicts classes in areas of the instance space not already predicted by other features.
- The **Fast Correlation-Based Filter** (FCBF) method [127] is a multivariate algorithm that measures feature-class and feature-feature correlation. FCBF starts by selecting a set of features that is highly correlated with the class based on symmetrical uncertainty (SU), which is defined as the ratio between the information gain and the entropy of two features. Then, it applies three heuristics that remove the redundant features and keep the features that are more relevant to the class. FCBF was designed for high-dimensionality data and has been shown to be effective in removing both irrelevant and redundant features. However, it fails to take into consideration the interaction between features.

- The **INTERACT** algorithm [128] uses the same goodness measure as the FCBF filter, i.e. SU, but it also includes the consistency contribution, which is an indicator of how significantly the elimination of a feature will affect consistency. The algorithm consists of two major parts. In the first part, the features are ranked in descending order based on their SU values. In the second part, features are evaluated one by one starting from the end of the ranked feature list. If the consistency contribution of a feature is less than an established threshold, the feature is removed, otherwise it is selected. The authors stated that this method can handle feature interaction, and efficiently selects relevant features.
- **Information Gain** [54] is one of the most common attribute evaluation methods. This univariate filter provides an ordered ranking of all the features and then a threshold is required. In this work the threshold will be set up selecting the features which obtain a positive information gain value.
- **ReliefF** [68] is an extension of the original Relief algorithm [67]. The original Relief works by randomly sampling an instance from the data and then locating its nearest neighbor from the same and opposite class. The values of the attributes of the nearest neighbors are compared to the sampled instance and used to update the relevance scores for each attribute. The rationale is that a useful attribute should differentiate between instances from different classes and have the same value for instances from the same class. ReliefF adds the ability of dealing with multiclass problems and is also more robust and capable of dealing with incomplete and noisy data. This method may be applied in all situations, has low bias, includes interaction among features and may capture local dependencies which other methods miss.
- The **mRMR** (minimum Redundancy Maximum Relevance) method [91] selects features that have the highest relevance with the target class and are also minimally redundant, i.e. it selects features that are maximally dissimilar to each other. Both optimization criteria (maximum-relevance and minimum-redundancy) are based on mutual information.

During the last five years, in addition to the application of known methods, an important number of filter approaches have been proposed and applied to microarray datasets, and this subsection will review the most interesting ones. An important number of filters are based on information theory, as can be seen in Section 3.1.1. On the other hand, several approaches include a preprocessing step to discretize data, since some filter require the data to be discrete, as reported in Section 3.1.2. Section 3.1.3 presents the filters which can deal with multiple binary problems and Section 3.1.4 describes methods which are related to other issues such as robustness. A summary of the up-to-date filters reviewed below can be found in Table 1.

3.1.1. Information theory

Firstly, we will present the methods based on information theory which, despite being introduced several years ago, are still the focus of much attention. A novel filter framework is presented in [122] to select optimal feature subsets based on a maximum weight and minimum redundancy (MWMR) criterion. The weight of each feature indicates its importance for some ad hoc tasks (e.g. clustering or classification) and the redundancy represents the correlation among features. With this method it is possible to select the feature subset in which the features are most beneficial to the subsequent tasks while the redundancy among them is minimal. Experimental results on five datasets (two of them based on the DNA microarray) demonstrated the advantage and efficiency of MWMR.

In [27] a statistical dependence measure is presented for gene selection in the context of DNA microarray classification. The proposed method is also based on a maximum relevance minimum redundancy approach, but it uses a simple measure of monotone dependence ($\mathscr{M}_d$) to quantify both relevance and redundancy. $\mathscr{M}_d$ was compared against the well-known minimum redundancy maximum relevance (mRMR) method, and was shown to obtain better or equal performance over binary datasets.

Also related to information theory, Meyer et al. [81] introduced MASSIVE, a new information-theoretic filter approach for mining microarray data. This filter relies on a criterion which consists of maximizing a term appearing both in the lower bound and the upper bound of the mutual information of a subset. The experimental results showed that the proposed method is competitive with five state-of-the-art approaches.

An entropic filtering algorithm (EFA) [48] was proposed as a fast feature selection method based on finding feature subsets that jointly maximize the normalized multivariate conditional entropy with respect to the classification ability of

Table 1
Filter methods used on microarray data. Type of evaluation (ranker/subset) and type of data (binary/multiclass).

Method	Original Ref.	Type (r/s)	Data (b/m)
BAHSIC	[107]	r	m
Discretizer + filter	[23,100]	s	m
EFA	[48]	s	b
$\mathscr{M}_d$	[27]	r	m
M_FS	[69]	r	m
MASSIVE	[81]	r	m
MWMR	[122]	s	b
PLS	[113]	r	m
RFS	[39]	r	m
RFS	[86]	r	m
RRFS	[39]	r	m

tumors. The solutions achieved are of comparable quality to previous results. They have been obtained in a maximum computing time of half an hour, using a very low number of genes.

In [107], the authors introduce a framework for feature selection based on dependence maximization between the selected features and the labels of an estimation problem, using the Hilbert–Schmidt Independence Criterion. Their proposed method, BAHSIC, is a filter method that demonstrated good performance on microarray data, compared with more specialized methods.

3.1.2. Discretization

After presenting the measures related to information theory, we will discuss the topic of discretization [44] related to feature selection. Although the use of a feature selection method when dealing with microarray data is a common practice, discretization has not received the same amount of attention. In [39] the authors proposed not only new techniques for feature selection, but also added a previous discretization step. They performed scalar feature discretization with the well-known Linde–Buzo–Gray algorithm, using a stopping criterion based on bit allocation. Then, the feature selection step applies a simple unsupervised criterion with indicators to the original numeric features and the discretized features. They also devised two efficient relevance/redundancy feature selection algorithms (RFS and RRFS) in order to remove redundant features.

In [23], the necessity of a previous discretization of the data is introduced for two main reasons: the first is to help the filtering process and the second is related to the high number of genes with very unbalanced values present in microarray datasets. The results of ten datasets demonstrated that the combination method, discretizer + filter, outperformed the results achieved by previous approaches, in some cases with improvements in the classification accuracy and a reduction in the number of genes needed.

3.1.3. Multiple binary problems

The same scheme of discretizer + filter was employed again in [100], but in this case to be applied only to multiclass datasets. While studies on feature selection using the multiclass approach (a method that can deal directly with multiple classes) are relatively frequent in the literature [27,81,107,23,39,86,69], very few studies employ the multiple binary sub-problems approach. Two well-known methods were employed for generating binary problems from a multiple class dataset: one versus one and one versus rest. The methodology was applied to 21 datasets, including a microarray dataset (Leukemia). With this dataset, the best results were obtained when applying feature selection. Specifically, the one versus rest approach obtained promising accuracy results along with a drastic reduction in the number of features needed.

Student and Fujarewicz [113] also proposed a method based on Partial Least Squares (PLS) and the decomposition of a set of two-class sub-problems; again using one versus one and one versus rest. They state that it is more effective to solve a multiclass feature selection by splitting it into a set of two-class problems and merging the results in one gene list. In this way, they obtained a very good accuracy rate and stability, as well as providing for the easy interpretation of the results by biologists.

3.1.4. Others

Robustness is a trending issue in feature selection. Nie et al. [86] proposed a new robust feature selection method (RFS) with emphasizing joint $\ell_{2,1}$-norm minimization in both the loss function and regularization. This method is robust to outliers and also efficient in calculation.

Finally, a very interesting and novel filter approach was proposed in [69] based on multi-task learning. When the number of labeled microarrays is particularly small (e.g. less than 10), the amount of available information diminishes to the level that even the most carefully designed classification approaches are bound to outperform. An alternative approach is to utilize information from the external microarray datasets, so accuracy in the target classification task can be significantly increased if data from the auxiliary tasks are consulted during learning. The multi-task filter (M_FS) was evaluated on microarray data showing that this method is successful when applied in conjunction with both single-task and multi-tasks classifiers.

3.2. Wrappers

As mentioned before, the wrapper approach has not received the same amount of attention as the filter methods, due to its high computational cost. As the number of features grows, the space of feature subsets grows exponentially. This is something that becomes a critical aspect when tens of thousands of features are considered. Furthermore, they have the risk of overfitting due to the small sample size of microarray data. As a result, the wrapper approach has been largely avoided in the literature.

Some works using the wrapper approach can be found in the earliest years of the investigation of microarray data. Notice that in a typical wrapper, a search is conducted in the space of genes. Then, the goodness of each gene subset found is evaluated by the estimated accuracy achieved by the specific classifier. This classifier is trained only with the found genes. For example, Inza et al. [61] evaluated classical wrapper search algorithms (sequential forward and backward selection, floating selection and best-first search) on three microarray datasets. Another example can be seen in [97], in which an incremental

wrapper called BIRS is presented for gene selection. Although the use of wrappers on microarray data has not evolved in the same line as the other feature selection methods, some examples were found in recent years.

Sharma et al. [102] propose an algorithm called successive feature selection (SFS). It is well-known that most of the conventional feature selection algorithms (e.g. individual ranking and forward selection schemes) have the drawback that a weakly ranked gene that could perform well in terms of classification with an appropriate subset of genes will be left out of the selection. Trying to overcome this shortcoming, the proposed SFS consists of first partitioning the features into smaller blocks. Once the top features from each of the blocks are obtained according to their classification performance, they are compared in order to obtain the best feature subset. This algorithm provides high classification accuracy on several DNA microarray datasets.

In [120], an evolutionary wrapper method (GA-KDE-Bayes) is presented. It uses a non-parametric density estimation method and a Bayesian classifier. The authors state that non-parametric methods are a good alternative for scarce and sparse data, such as the bioinformatics problem, since they do not make any assumptions about its structure and all the information comes from the data itself. Results from six microarray datasets showed a better performance than the others presented in the literature.

Table 2 visualizes the wrapper methods described, along with the original reference, the type of evaluation (ranker or subset) and the type of data that they can deal with (binary or multiclass).

3.3. Embedded

Despite its lower time consumption, a main disadvantage of the filter approach is the fact that it does not interact with the classifier, usually leading to worse performance results than those obtained with wrappers. However, we have seen that the wrapper model comes with an expensive computational cost, which is particularly aggravated by the high dimensionality of microarray data. An intermediate solution for researchers is the use of embedded methods, which use the core of the classifier to establish a criteria to rank features. Probably the most famous embedded method is Support Vector Machine based on Recursive Feature Elimination (SVM-RFE), proposed by Guyon et al. [51] to specifically deal with gene selection for cancer classification. This embedded method performs feature selection by iteratively training an SVM classifier with the current set of features and removing the least important feature indicated by the SVM. This method soon joined the group of algorithms which represent the state-of-the-art for gene selection, and therefore multiple extensions and modifications to it have been proposed. Next, we will describe several embedded approaches designed to deal with microarray data that we found when reviewing the up-to-date literature (for a summary of them, consult Table 3).

In [80] a new embedded method is introduced. It simultaneously selects relevant features during classifier construction by penalizing each feature's use in the dual formulation of support vector machines (SVM). This approach is called kernel-penalized SVM (KP-SVM) and it optimizes the shape of an anisotropic RBF Kernel eliminating features that have low relevance for the classifier. The experiments on two benchmark microarray datasets and two real-world datasets showed that KP-SVM outperformed the alternative approaches and determined consistently fewer relevant features.

Wang et al. [119] proposed a First Order Inductive Learner (FOIL) rule based feature subset selection algorithm, called FRFS. This method first generates the FOIL classification rules using a modified propositional implementation of the FOIL algorithm. Then, it combines the features that appeared in the antecedents of all of the rules, and achieves a candidate feature subset that excludes redundant features and reserves the interactive ones. Lastly, it measures the relevance of the features in the candidate feature subset by their proposed new metric CoverRatio and identifies and removes the irrelevant features.

Table 2
Wrapper methods used on microarray data. Type of evaluation (ranker/subset) and type of data (binary/multiclass).

Method	Original Ref.	Type (r/s)	Data (b/m)
GA-KDE-Bayes	[120]	s	b
SPS	[102]	s	m

Table 3
Embedded methods used on microarray data. Type of evaluation (ranker/subset) and type of data (binary/multiclass).

Method	Original Ref.	Type (r/s)	Data (b/m)
FRFS	[119]	s	m
IFP	[31]	s	b
KP-SVM	[80]	s	m
PAC-Bayes	[101]	r	b
Random Forest	[16]	s	m

Shah et al. [101] not only focus on obtaining a small number of genes but also on having verifiable future performance guarantees. They investigated the premise of learning conjunctions of decision stumps and proposed three formulations based on different learning principles, which embed the feature selection as a part of the learning process itself. One of their proposals, Probably Approximately Correct (PAC) Bayes, yields competitive classification performance while at the same time utilizing significantly fewer attributes.

In [31], the iterative perturbation method (IFP), an embedded gene selector, is introduced and applied to four microarray datasets. This algorithm uses a backward elimination approach and a criterion to determine which features are the least important, which relies on the classification performance impact that each feature has when perturbed by noise. If adding noise leads to a big change in the classification performance, then the feature is considered relevant. The IFP approach resulted in comparable or superior average class accuracy when compared to well-studied SVM-RFE on three out of the four datasets.

To overcome the problem of the imbalance of some microarray datasets, a new embedded method based on the random forest algorithm is presented in [16]. Its strategy is composed of different methods and algorithms. First, an algorithm to find the best training error cost for each class is run, in order to deal with the imbalance of the data. Then, random forest is run to select the relevant features. Finally, a strategy to avoid overfitting is also applied. The method was designed ad hoc to deal with a complex gene expression dataset for Leukaemia malignancy, showing a very acceptable outcome.

3.4. Other algorithms

Nowadays, the trend is to use not only classical feature selection methods (filters, wrappers and embedded) but also to focus on new combinations such as hybrid or ensemble methods.

Hybrid methods usually combine two or more feature selection algorithms of different conceptual origin in a sequential manner. In [85] two of the most famous feature selection methods for microarray data are combined: SVM-RFE and mRMR. They propose an approach that incorporates a mutual-information-based mRMR filter in SVM-RFE to minimize the redundancy among selected genes. Their approach improved the accuracy of classification and yielded smaller gene sets compared with mRMR and SVM-RFE, as well as other popular methods.

Shreem et al. [105] also used mRMR in their hybrid method. In this case, the proposed approach combines ReliefF, mRMR and GA (Genetic Algorithm) coded as R-m-GA. In the first stage, the candidate gene set is identified by applying ReliefF. Then, the redundancy is minimized with the help of mRMR, which facilitates the selection of effectual gene subsets from the candidate set. In the third stage, GA with classifier (used as a fitness function by the GA) is applied in order to choose the most discriminating genes. The proposed method is capable of finding the smallest gene subset that offers the highest classification accuracy.

Chuang et al. [33] proposed a hybrid method called CFS-TGA, which combines correlation-based feature selection (CFS) and the Taguchi-genetic algorithm, in which the K-nearest neighbor served as a classifier. The proposed method obtained the highest classification accuracy in ten out the eleven gene expression datasets that it was tested on.

In [71] another hybrid method is proposed. It first uses a genetic algorithm with dynamic parameter setting (GADP) to generate a number of subsets of genes and to rank the genes according to their occurrence frequencies in the gene subsets. Then, χ^2 is used to select a proper number of the top-ranked genes for data analysis. Finally, an SVM is employed to verify the efficiency of the selected-genes. The experimental results on six microarray datasets showed that the GADP method is better than the existing methods in terms of the number of selected genes and the prediction accuracy.

Leung and Hung [73] proposed a multiple-filter-multiple-wrapper (MFMW) method. The rationale behind this proposal is that filters are fast but their predictions are inaccurate whilst wrappers maximize classification accuracy at the expense of a formidable computation burden. MFMW is based on previous hybrid approaches that maximize the classification accuracy for a chosen classifier with respect to a filtered set of genes. The drawback of the previous hybrid methods which combine filters and wrappers is that classification accuracy is dependent on the choice of a specific filter and wrapper. MFMW overcomes this problem by making use of multiple filters and multiple wrappers to improve the accuracy and robustness of the classification.

Ensemble feature selection builds on the assumption that combining the output of multiple experts is better than the output of any single expert. Typically, ensemble learning has been applied to classification, but it has recently been applied to microarray gene selection. An ensemble of filters (EF) is proposed in [24]. The rationale of this approach is behind the variability of results of each available filter over different microarray datasets. That is, a filter may obtain excellent classification results in a given dataset while performing poorly in another dataset, even if it is in the same domain. This ensemble obtains a classification prediction for every different filter forming the ensemble, and then combines these predictions by simple voting. Experiments on 10 microarray datasets show that the ensemble obtained the lowest average classification error for the four classifiers tested. Recently, the same authors introduced new ensembles to improve performance (E1-cp, E1-ni, E1-ns, E2) [25].

From the perspective of pattern analysis, researchers must focus not only on classification accuracy but also on producing stable or robust solutions. Trying to improve the robustness of feature selection algorithms, Yang and Mao [124] proposed an ensemble method called multicriterion fusion-based recursive feature elimination (MCF-RFE). Experimental studies on microarray datasets demonstrated that the MCF-RFE method outperformed the commonly used benchmark feature selection algorithm SVM-RFE both in classification performance and in terms of the stability of feature selection results.

Abeel et al. [11] are also concerned with the analysis of the robustness of biomarker selection techniques. They proposed a general experimental setup for stability analysis that can easily be included in any biomarker identification pipeline. In addition, they also presented a set of ensemble feature selection methods improving biomarker stability and classification performance in four microarray datasets. They used recursive feature elimination (RFE) as a baseline method and a bootstrapping method to generate diversity in the selection. Then, two different schemes were proposed to aggregate the different rankings of features. They found that when the number of selected features decreases, the stability of RFE tends to degrade while ensemble methods offer significantly better stability.

Ye et al. [125] proposed a stratified sampling method to select the feature subspaces for random forest (SRF). The key idea is to stratify features into two groups. One group will contain strong informative features and the other weak informative features. Then, for feature subset selection, features are randomly selected from each group proportionally. The advantage of stratified sampling is that it can ensure that each subspace contains enough informative features for classification in high dimensional data.

Clustering methods for microarray data have been also recently proposed. Most of the gene selection techniques are based on the assumption of the independence between genes (actually a typical approach is to rank the genes individually). However, it is well known that genes interact with each other through gene regulative networks. To overcome this problem, Lovato et al. [78] presented a novel feature selection scheme, based on the Counting Grid (GC) model [64], which can measure and consider the relation and influence between genes.

Song et al. [108] presented a fast clustering-based feature selection algorithm (FAST) which works in two steps. In the first step, features are divided into clusters by using graph-theoretic clustering methods. In the second step, the most representative feature that is strongly related to target classes is selected from each cluster to form a subset of features. Since features in different clusters are relatively independent, the clustering-based strategy of FAST has a high probability of producing a subset of useful and independent features. The exhaustive evaluation carried out on 35 real-world datasets (14 of them in the microarray domain) demonstrated that FAST not only produced smaller subsets of features, but also improved the performances of four types of classifiers.

Table 4 depicts a summary of the methods presented in this section. The original reference is displayed, as well as the type of evaluation (ranker or subset) and the type of data they can deal with (binary or multiclass).

4. Microarray datasets

After reviewing the most up-to-date feature selection methods dealing with microarray data, this section will be focused on the particularities of the datasets. First, Section 4.1 will enumerate the existing microarray data repositories, whilst Section 4.2 provides a summary of the characteristics of the most famous binary and multiclass datasets used in the literature. Finally, Section 4.3 is devoted to an analysis of the problematics of microarray data, such as the small-sample size, the imbalance of the data, the dataset shift or the presence of outliers.

4.1. DNA microarray repositories

Although in the initial development of DNA microarray data analysis it was difficult to find datasets to deal with, in recent years there has been a growing number of public microarray data repositories of a wide spectrum of cancer types available for the scientific community. The most famous are listed below:

- *ArrayExpress*, from the European Bioinformatics Institute [1]:
  http://www.ebi.ac.uk/arrayexpress/

Table 4
Other feature selection methods used on microarray data. Type of evaluation (ranker/subset) and type of data (binary/multiclass).

Method	Original Ref.	Type (r/s)	Data (b/m)
CFS-TGA	[33]	s	m
E1-cp	[25]	s	b
E1-ni	[25]	s	b
E1-ns	[25]	s	b
E2	[25]	s	b
Ensemble RFE	[11]	s	b
EF	[24]	s	m
FAST	[108]	s	m
GADP	[71]	s	m
GC	[78]	r	b
MCF-RFE	[124]	s	b
MFMW	[73]	s	b
R-m-GA	[105]	s	b
SRF	[125]	s	m
SVM-RFE with MRMR	[85]	r	b

- *Cancer Program Data Sets*, from the Broad Institute [2]: http://www.broadinstitute.org/cgi-bin/cancer/datasets.cgi
- *Dataset Repository*, from the Bioinformatics Research Group of Universidad Pablo de Olavide [3]: http://www.upo.es/eps/bigs/datasets.html
- *Feature Selection Datasets*, from Arizona State University [5]: http://featureselection.asu.edu/datasets.php
- *Gene Expression Model Selector*, from Vanderbilt University [110]: http://www.gems-system.org
- *Gene Expression Omnibus*, from the National Institutes of Health [6]: http://www.ncbi.nlm.nih.gov/geo/
- *Gene Expression Project*, from Princeton University [7]: http://genomics-pubs.princeton.edu/oncology/
- *Kent Ridge Bio-Medical Dataset Repository*, from the Agency for Sciency, Technology and Research [8]: http://datam.i2r.a-star.edu.sg/datasets/krbd
- *Stanford Microarray Database*, from Stanford University [10]: http://smd.stanford.edu/

4.2. Datasets

As mentioned in the Introduction, there are two types of microarray datasets present in the literature. The most famous are the binary datasets, usually related to separating healthy patients from cancer patients. When the goal is to distinguish between different types of tumors, we can find multiclass datasets, in which the classification task becomes more complicated.

Table 5 displays the binary microarray datasets used in the research mentioned in Section 3. There the number of samples *s*, the number of features *f*, the class distribution, the original reference of the dataset, the year when the dataset was published, the references of the works using the dataset and where it is available for download, can all be found. When a data is not available, it is represented as "*unknown*". In turn, Table 6 visualizes the multiclass microarray datasets. In this case, *c* stands for the number of classes and the class distribution is not shown due to the high diversity in the number of classes.

4.3. Intrinsic characteristics of microarray data

As mentioned in the Introduction, microarray data classification poses a serious challenge for computational techniques, because of their large dimensionality (up to several tens of thousands of genes) with small sample sizes. Furthermore, there are additional experimental complications that render the analysis of microarray data an intriguing domain.

4.3.1. Small sample size

The first problem that one may find when dealing with microarray data is related to the small sample size (usually less than 100). A key point in this regard is that error estimation is greatly impacted by small samples [35]. Without the appropriate estimation of the error, an unsound application of classification methods follows, which has generated a large number of publications and an equally large amount of unsubstantiated scientific hypotheses [28]. For example, in [82] it is reported that reanalysis of data from the seven largest published microarray-based studies that have attempted to predict the prognosis of cancer patients reveals that five of those seven did not classify patients better than chance. To overcome this problem, it becomes necessary to select a correct validation method for estimating the classification error. This issue will be further discussed in Section 5.1.

4.3.2. Class imbalance

A common problem in microarray data is the so-called *class imbalance problem*. This occurs when a dataset is dominated by a major class or classes which have significantly more instances than the other rare/minority classes in the data [56,116,76]. Typically, people have more interest in learning rare classes. For example, in the domain at hand, the cancer class tends to be rarer than the non-cancer class because usually there are more healthy patients. However, it is important for practitioners to predict and prevent the appearance of cancer. In these cases, standard classifier learning algorithms have a bias toward the classes with a greater number of instances, since the rules that correctly predict those instances are positively weighted in favor of the accuracy metric, whereas specific rules that predict examples from the minority class are usually ignored (treated as noise), because more general rules are preferred. Therefore, minority class instances are more often misclassified than those from the other classes [41].

Although class imbalance does not hinder the learning task by itself, there are some difficulties related to this problem that occur, such as a small sample size, as is the case with microarray data. Examples of very unbalanced microarray datasets are Lung_test, Prostate_test or CNS, among others (see Table 5). This problematic is of special importance when the imbalance is more marked in the test set than in the training set, as will be further discussed in Section 4.3.4, which deals with the dataset shift problem. Multiclass datasets also suffer from this problem. For example, Lymphoma dataset (see Table 6) has 9 classes but the majority class takes 48% of the samples.

Table 5
Dataset description for binary datasets: s and f are the number of samples and features, respectively.

Dataset	s	f	Distribution	Original Ref.	Year	Where used	Download
B_MD	34	7129	26–74%	[95]	2002	[101]	*unknown*
Bone Lesion	173	12,625	*unknown*	[117]	2003	[101]	*unknown*
Brain	21	12,625	33–67%	[88]	2003	[27,120]	[2]
Brain_Tumor1	60	7129	*unknown*	*unknown*	*unknown*	[81]	*unknown*
Brain_Tumor2	50	12,625	*unknown*	*unknown*	*unknown*	[81]	*unknown*
Breast	22	3226	*unknown*	[58]	2001	[71]	*unknown*
Breast Cancer	97	24,481	*unknown*	[118]	2002	[24,23,27]	[8]
Breast-test	19	24,481	37–63%	[118]	2002	[23,48]	[8]
Breast-train	78	24,481	56–44%	[118]	2002	[23,48]	[8]
BreastER	49	7129	49–51%	[123]	2001	[101]	*unknown*
BR-ER49	49	6817	49–51%	[123]	2001	[73]	*unknown*
C_MD	60	7129	35–65%	[95]	2002	[101]	*unknown*
Celiac	132	22,185	*unknown*	[57]	2009	[101]	*unknown*
CNS/Embrional-T	60	7129	35–75%	[95]	2002	[124,24,25,23,27,105,125]	[2,3]
Colon	62	2000	35–65%	[14]	1999	[39,101,78,107,80,71,124,11,24,23,27,73,85,48,122,108,31,125,120]	[2,7,8,3]
Colon-epi	202	44,290	*unknown*	[87]	2008	[101]	*unknown*
DLBCL	77	5470	75–25%	[104]	2002	[39,33,122,125]	[110]
DLBCL	47	4026	49–51%	[13]	2000	[71,24,25,23,27,105]	[8,2]
DLBCL	77	7129	75–25%	[104]	2002	[124,81]	[2,9]
GLI-85	85	22,283	31–69%	[40]	2004	[27]	[5]
Leukemia/ ALLAML	72	7129	35–65%	[46]	1999	[39,101,107,71,124,11,86,24,25,27,73,85,31,120]	[2,8]
Leukemia_test	34	7129	71–29%	[46]	1999	[119,23,48,108]	[3,2,8]
Leukemia_train	38	7129	59–41%	[46]	1999	[119,23,48,108]	[3,2,8]
Lung	52	918	75–25%	[43]	2001	[101]	*unknown*
Lung	181	12,533	83–17%	[49]	2002	[24,25,27,48]	[8,2]
Lung	410	2428	34–66%	[103]	2008	[31]	*unknown*
Lung_test	149	12,533	90–10%	[49]	2002	[23,48]	[8,2]
Lung_train	32	12,533	50–50%	[49]	2002	[23,48]	[8,2]
LUNG181	181	12,600	17–83%	[49]	2002	[73]	*unknown*
LYM77	77	6817	25–75%	[104]	2002	[73]	[8]
Lymphoma/B-cell1	45	4026	49–51%	[13]	2000	[11,108]	[2]
Moffitt colon cancer	122	2619	31–69%	[38]	2005	[31]	*unknown*
Ovarian	253	15,154	36–64%	[93]	2002	[24,25,23,27,120,107]	[8]
Prostate	102	6033	51–49%	[106]	2002	[78,11]	*unknown*
Prostate	136	12,600	43–57%	[106]	2002	[124,24,25,27]	[2,8]
Prostate-test	34	12,600	26–74%	[106]	2002	[102,85,23,48]	[8]
Prostate-train	102	12,600	49–51%	[106]	2002	[102,86,73,85,81,23,48,105]	[8,9]
Prostate Tumor	102	10,509	51–49%	[106]	2002	[39,33,125,120]	[110]
SMK-CAN-187	187	19,993	48–52%	[109]	2007	[27,119,108]	[5]

The traditional preprocessing techniques used to overcome this issue are undersampling methods, which create a subset of the original dataset by eliminating instances; oversampling methods, which create a superset of the original dataset by replicating some instances or creating new instances from existing ones; and finally, hybrid methods that combine both sampling methods. One of the most employed resampling techniques is the so-called SMOTE [32], in which the minority class is over-sampled by taking each minority class sample and introducing synthetic examples along the line segments joining any/all of the k minority class nearest neighbors. This technique was applied in [21] on microarray data, although the authors stated that it does not attenuate the bias towards classification in the majority class for most classifiers. In recent years, ensemble of classifiers has arisen as a possible solution to the class imbalance problem, attracting great interest among researchers [41,42], in several cases combined with preprocessing techniques such as SMOTE.

Ensemble-based algorithms have been proven to improve the results that are obtained by the usage of data preprocessing techniques and training a single classifier [41]. For all these reasons, it is worth considering this problematic when dealing with unbalanced microarray datasets.

4.3.3. Data complexity

Data complexity measures are a recent proposal to represent characteristics of the data which are considered difficult in classification tasks, such as the overlapping among classes, their separability or the linearity of the decision boundaries [99,59]. These measures have been applied particularly to gene expression analysis in [77,89], demonstrating that low complexity corresponds to small classification error. In particular, the measures of *class overlapping*, such as F1 (*maximum Fisher's discriminant ratio*) [99], focus on the effectiveness of a single feature dimension in separating the classes. They examine the range and spread of values in the dataset within each class and check for overlapping among different classes.

Table 6
Dataset description for multiclass datasets: *s*, *f* and *c* are the number of samples, features and classes, respectively.

Dataset	*s*	*f*	*c*	Original Ref.	Year	Where used	Download
9-Tumors	60	5726	9	[111]	2001	[39,33,81,125]	[110,9]
11-Tumors/Carcinomas	174	12,533	11	[114]	2001	[39,33,86,81,125]	[110,9]
14-Tumors	308	15,009	26	[96]	2001	[39,33,69,81,125]	[110,9]
Brain Tumor 1	90	5920	5	[95]	2002	[39,33,81,125]	[110,9]
Brain Tumor 2	50	10,367	4	[88]	2003	[39,33,81]	[110,9]
CLL-SUB-111	111	11,340	3	[55]	2004	[119,108]	[5]
GCM	198	16,306	14	[96]	2001	[71,24,23,27]	[2]
GCM	190	16,063	14	[96]	2001	[125]	[3]
GLA-BRA-180	180	49,151	4	[115]	2006	[27,108]	[5]
Glioma	50	12,625	4	[88]	2003	[86]	[2]
Global Cancer Map/GCM-Train	144	16,063	14	[96]	2001	[119,108]	[3]
Leukaemia	110	22,278	*unknown*	*unknown*	*unknown*	[16]	*unknown*
Leukemia 1	72	5327	3	[46]	1999	[39,33,81,125]	[110,9]
Leukemia 2	72	11,225	3	[17]	2002	[39,33,81,125]	[110,9]
Lung	254	8359	5	[20]	2001	[113]	[2]
Lung-Cancer	203	12,601	5	[20]	2001	[39,33,86,81,125]	[110,9]
Lymphoma/B-cell3	96	4026	9	[13]	2000	[39,80,24,23,27,108,120]	[110]
MLL	72	8359	3	[17]	2002	[113]	[2]
MLL-train	57	12,582	3	[17]	2002	[102]	[8]
MLL-test	15	12,582	3	[17]	2002	[102]	[8]
TOX-171	171	5748	4	[112]	2010	[27,119,108]	[5]
SRBCT	83	2309	4	[66]	2001	[39,113,33,71,81]	[110,9]
SRBCT-train	63	2309	4	[66]	2001	[102]	[110]
SRBCT-test	20	2309	4	[66]	2001	[102]	[110]
Yeast	79	2467	*unknown*	[30]	2000	[107]	*unknown*

4.3.4. Dataset shift

Another common problem when datasets were originally divided to training and test sets, is the so-called *dataset shift*. This occurs when the testing (unseen) data experience a phenomenon that leads to a change in the distribution of a single feature, a combination of features, or the class boundaries [83]. As a result, the common assumption that the training and testing data follow the same distributions is often violated in real-world applications and scenarios, which may hinder the process of feature selection and classification. For example, Lung, Leukemia and Prostate datasets have separated training and test sets (see Table 5). In the case of Lung, there is a single feature (#1136) which can correctly classify all the samples in the training set, as shown in Fig. 2(a), in which different colors and shapes stand for different classes and the dashed line shows a clear linear separation between them. However, the same feature is not that informative in the test set and the class is not linearly separable, as displayed in Fig. 2(b). Furthermore, note that there is an enormous disparity in class distribution: 50–50% in the training set and 90–10% in the test set.

The Prostate dataset poses a big challenge for machine learning methods since the test dataset was extracted from a different experiment and has a nearly 10-fold difference in overall microarray intensity from the training data. In fact, the test distribution (26–74%) differs significantly from the train distribution (49–51%) and with an inappropriate feature selection, some classifiers simply assign all the samples to one of the classes [23,25].

Dataset shift can also appear when performing a cross-validation technique, which divides the whole training set into several subsets of data. In this case, there are some partitioning methods which may solve the problem [84].

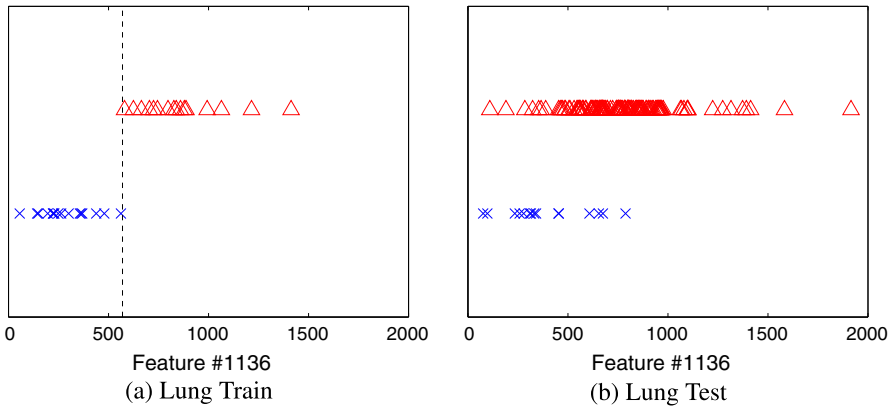

Fig. 2. Feature #1136 in Lung dataset.

4.3.5. Outliers

An important aspect that has been neglected in the literature is to detect outliers [18] in microarray samples. In some microarray datasets, there are samples that are incorrectly labeled or identified as likely to be contaminated which should in fact be designated outliers, since they can exert a negative effect on the selection of informative genes for sample classification. In [65], a method was developed which found some outlying samples in the well-known Colon dataset. Therefore, analysis of samples designated as outliers should be considered as a pre-processing step in the classification of microarray datasets because they can have a negative effect on the gene subset selection and, as a consequence, on the final prediction [47].

5. An experimental study in binary classification: analysis of results

The goal of performing feature selection on microarray data can be twofold: class prediction or biomarkers' identification. If the goal is class prediction, there is a demand for machine learning techniques such as supervised classification. However, if the objective is to find informative genes, the classification performance is ignored and the selected genes have to be individually evaluated. The experiments that will be presented in this section are focused on class prediction, which is an important reason to use feature selection methods in microarray analysis. The typical microarray pipeline is formed by a feature selection step, followed by a classification stage which provides an error estimation, as seen in Fig. 3.

The rest of this section is devoted to the importance of the validation techniques usually applied on microarray data and to analyzing the characteristics of the datasets whilst providing classification accuracy results obtained with classical feature selection methods in order to help the readers compare the performance of their novel methods on microarray data. Finally, an analysis of the results will be presented.

5.1. Validation techniques

To evaluate the goodness of the selected set of genes, it is necessary to have an independent test set with data which have not been seen by the feature selection method. In some cases, the data has been originally distributed into training and test sets, so the training set is usually employed to perform the feature selection process and the test set is used to evaluate the appropriateness of the selection. However, not all the datasets found in the literature are originally partitioned. In order to overcome this issue, several validation techniques exist, and in what follows we describe the most widely used ones in the microarray domain.

One of the most widely used validation techniques in the microarray domain is the so-called *k-fold cross validation*, along with its variant *leave-one-out cross validation*. However, it has been shown [84] that cross-validation can potentially introduce dataset shift (see Section 4.3.4), a harmful factor that is often not taken into account and which can result in inaccurate performance estimation. To solve this problem, *Distribution optimally balanced stratified cross-validation* (DOB-SCV) [84] is based on the idea that, by assigning close-by examples to different folds, each fold will end up with enough representatives of every region, thus avoiding dataset shift. To achieve this goal, DOB-SCV starts on a random unassigned example, and then finds its $k-1$ nearest unassigned neighbors of the same class. Once it has found them, it assigns each of those examples to a different fold. The process is repeated until all examples have been assigned. Other popular validation techniques are the well-known *bootstrap* resampling strategy [37] and the *holdout validation*.

The selection of a validation technique for its use in the microarray domain is not an easy-to-solve question. This is due to the fact that microarray data is characterized by an extremely high number of features and a comparatively small number of samples. This situation is commonly referred to as a *small-sample* scenario, which means that the application of traditional pattern recognition methods must be carried out with careful judgment in order to avoid pitfalls [28].

A key point for microarray classification is that error estimation is greatly impacted by small samples [35], so the choice of a validation technique must be further discussed. In fact, there are several works in the literature dealing with this issue. Ambroise and McLachlan [15] recommend the use of 10-fold cross validation rather than leave-one-out and, concerning the bootstrap, they suggest using the so called 0.632+ bootstrap error estimate designed to handle overfitted prediction rules. In [29], an extensive simulation study was performed comparing cross-validation, resubstitution and bootstrap estimation. They state that, while cross-validation error estimation is much less biased than resubstitution, it displays excessive variance, which makes individual estimates unreliable for small samples. Bootstrap methods provide improved performance relative to variance, but at a high computational cost and often with increased bias.

In this situation, a best validation technique for microarray data does not exist. In fact, reviewing the recent literature one can find examples of the four methods described above. *k*-fold cross-validation is a common choice [81,107,86,101,31,105,125], as well as holdout validation [48,39,102,80,16,71,78,75]. Bootstrap sampling was less used [113,85,124], probably due to its high computational cost, and there are also some representatives of leave-one-out cross-validation [33,73,92].

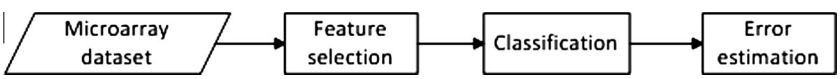

Fig. 3. DNA microarray classification pipeline.

5.2. On the datasets' characteristics

In this study, we have considered nine widely-used binary microarray datasets, which are available for download in [5,110,8]. The reason for choosing binary datasets is that they are much more common in the literature than the multiclass ones. As a matter of fact, a typical microarray dataset consists of distinguishing between having a given cancer or not, therefore the great majority of the datasets are binary. Tables 7 and 8 summarize the properties of the selected datasets: for each dataset, the number of features (# Feats.), number of samples (# Samp.) and the percentage of examples of each class is shown. The imbalance ratio [90] (IR) is defined as the number of negative class samples divided by the number of positive class samples, in which a high level indicates that the dataset is highly imbalanced. Finally, F1 (*maximum Fisher's discriminant ratio*, see Section 4.3.3) [59] checks for overlapping among the classes in which the higher the F1, the more separable the data is. Notice that for the datasets in Table 7, which were originally divided to training and test sets, these measures are shown for both partitions. It is shown that both of these datasets present more imbalance in the test set than in the training set, especially in the case of the Prostate dataset. As for the F1 measure, the higher amount of overlapping occurs on the Breast training set, with a value of 0.68. Regarding the information depicted in Table 8, the most unbalanced dataset is GLI and the one most affected by overlapping is SMK, with an F1 value of 0.41.

5.3. Feature selection methods

Seven classical feature selection methods widely used by the researchers in this field were chosen to be applied in this study (CFS, FCBF, INTERACT, Information Gain, ReliefF, mRMR and SVM-RFE). All of them are available in the well-known Weka tool [53], except for the mRMR filter, whose implementation is available for Matlab ®. Although there are other accessible feature selection methods implemented in popular tools such as KMINE [19], KEEL [12] or *Feature Selection Algorithms* from Arizona State University [4], we have opted for these seven methods because they are used extensively in the literature [70,122,107,31,81]. A brief description for each of them can be found in Section 3. Their performance may serve as a reference for the interested reader, so that comparative studies can easily be carried out based on this material.

5.4. Evaluation measures

In order to evaluate the behavior of the feature selection methods after applying a classifier, three well-known measures are used: accuracy, sensitivity and specificity. These metrics were designed to measure the performance of a binary classification test. In general, sensitivity indicates how well the test predicts the actual positives (e.g. the percentage of cancer patients who are correctly identified as having the condition) while specificity measures how well the test identifies the negatives (e.g. the percentage of healthy patients who are correctly identified as not having cancer). A perfect predictor would be described as 100% sensitive (e.g. predicting all patients with cancer as such) and 100% specific (e.g. not predicting any patient from the healthy group as having cancer). Accuracy is expected to measure how well the test predicts both categories.

Sensitivity, specificity and accuracy can be described in terms of true positives (TP), true negatives (TN), false negatives (FN) and false positives (FP):

- $Sensitivity = \frac{TP}{TP+FN}$
- $Specificity = \frac{TN}{TN+FP}$
- $Accuracy = \frac{TN+TP}{TN+TP+FN+FP}$

Table 7
Summary description of the binary datasets used in the holdout experimental study.

Dataset	# Feats.	Train				Test			
		# Samp.	(%min,%maj)	IR	F1	# Ex.	(%min,%maj)	IR	F1
Breast	24,481	78	(43.59,56.41)	1.29	0.68	19	(36.84,63.16)	1.71	4.98
Prostate	12,600	102	(49.02,50.98)	1.04	2.05	34	(26.47,73.53)	2.78	11.35

Table 8
Summary description of the binary datasets used in the *k*-fold cross-validation experimental study.

Dataset	# Feats.	# Samp.	(%min,%maj)	IR	F1
Brain	12,625	21	(33.33,66.67)	2.00	0.89
CNS	7129	60	(35.00,65.00)	1.86	0.45
Colon	2000	62	(35.48,64.52)	1.82	1.08
DLBCL	4026	47	(48.94,51.06)	1.04	2.91
GLI	22,283	85	(30.59,69.41)	2.27	2.35
Ovarian	15,154	253	(35.97,64.03)	1.78	6.94
SMK	19,993	187	(48.13,51.87)	1.08	0.41

5.5. Analysis of results

The goal of this section is to perform an experimental study using the most representative binary datasets and some classical widely used feature selection methods, providing the readers with some baseline for their comparisons. To evaluate the adequacy of these methods over microarray data, three well-known classifiers were chosen: C4.5, naive Bayes and SVM (Support Vector Machine). As reported in Section 5.1, there is no consensus in the literature about which validation technique to use when dealing with microarray data. In the light of these facts, the authors have opted to perform two studies. In the first one, a holdout validation will be applied to those datasets which were originally divided to training and test datasets. As revealed in Section 4.3.4, the training and test data of these datasets were extracted under different conditions, which presents an added challenge for the machine learning methods. If the two sets are joined in a unique dataset (e.g. in order to later apply a *k*-fold cross-validation), the new situation would be easier for the learner, and this particular characteristic of microarray data would be overlooked. The second study will consist of applying a 5-fold cross-validation over those datasets which provide a unique training test, in which 5 folds have been chosen, because, with the standard value of 10, in some datasets the test set would remain with only a couple of samples. However, as mentioned in Section 5.1, in some cases cross-validation can potentially introduce dataset shift, so we will include DOB-SCV in the experimental study in an attempt to overcome this problem.

The three first feature selection methods (CFS, FCBF and INTERACT) return a subset of features, whilst the remaining four (IG, ReliefF, mRMR and SVM-RFE) provide an ordered ranking of the features. For the ranker methods, we show the performance when the top 10 and top 50 features are retained. For those methods which return a subset of features, the number of features selected for each dataset is shown in Table 9. Notice that, for the datasets involved in the cross-validation study (Brain, CNS, Colon, DLBCL, Gli85, Ovarian and Smk), this number is the mean average of the number of features selected in each fold. Since we are testing two types of partitions, both values are shown in the table (regular cross-validation/ DOB-SCV).

5.6. Holdout validation study

This section reports the experimental results achieved over the binary datasets that were originally divided into training and test sets (see Table 7). Tables 10–12 show the results achieved by C4.5, naive Bayes and SVM, respectively. These tables depict the classification accuracy (Ac), sensitivity (Se) and specificity (Sp) of the test datasets. For the sake of comparison, the first row shows those values without applying feature selection techniques. Notice that the C4.5 algorithm carries out a feature selection because not all the attributes are considered when constructing the tree. The best results for the dataset and classifier are marked in bold face.

Analyzing these tables, it is notable that the results obtained with SVM considerably outperformed those achieved by C4.5 or naive Bayes. In fact, there is a clear tendency in the literature to use SVM as the standard *de facto* method to estimate performance measures and in [47], it is stated that the superiority of SVMs to other several classifiers seems to be beyond doubt. As mentioned in Section 4.3.4, the Prostate dataset suffers from dataset shift, since the test dataset was extracted from a different experiment, and apparently C4.5 and naive Bayes classifiers cannot solve this problem and opted to assign all the examples to the majority class.

Table 9
Number of features selected by subset methods on binary datasets.

Method	Brain	Breast	CNS	Colon	DLBCL	Gli85	Ovarian	Prostate	Smk
CFS	36/49	130	44/44	24/25	61/65	141/156	35/33	89	107/103
FCBF	1/1	99	33/35	14/15	35/37	116/118	27/26	77	50/55
INT	36/49	102	33/34	14/16	45/51	117/123	32/31	73	51/51

Table 10
Experimental results for C4.5 classifier on binary datasets after performing holdout validation.

		Breast			Prostate		
		Ac	Se	Sp	Ac	Se	Sp
no FS		0.74	**1.00**	0.58	0.26	**1.00**	0.00
CFS		0.68	0.71	0.66	0.26	**1.00**	0.00
FCBF		0.58	0.28	0.75	0.26	**1.00**	0.00
INT		**0.79**	0.71	0.83	0.26	**1.00**	0.00
IG	#10	0.47	0.28	0.58	0.26	**1.00**	0.00
	#50	0.53	0.42	0.58	0.29	**1.00**	0.04
ReliefF	#10	0.58	0.28	0.75	0.26	**1.00**	0.00
	#50	0.42	0.71	0.25	0.29	**1.00**	0.04
SVM-RFE	#10	0.58	**1.00**	0.33	0.32	**1.00**	0.08
	#50	0.58	**1.00**	0.33	0.26	**1.00**	0.00
mRMR	#10	0.58	0.71	0.50	**0.41**	0.88	**0.24**
	#50	0.74	0.42	**0.91**	0.35	**1.00**	0.12

Table 11
Experimental results for naive Bayes classifier on binary datasets after performing holdout validation.

		Breast			Prostate		
		Ac	Se	Sp	Ac	Se	Sp
no FS		0.37	**1.00**	0.00	**0.26**	**1.00**	0.00
CFS		0.37	**1.00**	0.00	**0.26**	**1.00**	0.00
FCBF		0.37	**1.00**	0.00	**0.26**	**1.00**	0.00
INT		0.37	**1.00**	0.00	**0.26**	**1.00**	0.00
IG	#10	0.32	0.85	0.00	**0.26**	0.88	0.04
	#50	0.37	**1.00**	0.00	0.24	0.88	0.00
ReliefF	#10	0.74	0.71	0.75	0.21	0.55	**0.08**
	#50	**0.89**	0.85	**0.91**	0.21	0.77	0.00
SVM-RFE	#10	0.68	0.85	0.58	0.18	0.55	0.04
	#50	0.63	**1.00**	0.41	**0.26**	**1.00**	0.00
mRMR	#10	0.37	**1.00**	0.00	**0.26**	**1.00**	0.00
	#50	0.37	**1.00**	0.00	**0.26**	**1.00**	0.00

Table 12
Experimental results for SVM classifier on binary datasets after performing holdout validation.

		Breast			Prostate		
		Ac	Se	Sp	Ac	Se	Sp
no FS		0.58	0.85	0.41	0.53	**1.00**	0.36
CFS		0.68	0.85	0.58	**0.97**	**1.00**	**0.96**
FCBF		0.58	0.28	0.75	**0.97**	**1.00**	**0.96**
INT		0.74	0.71	0.75	0.71	**1.00**	0.60
IG	#10	0.58	0.71	0.50	**0.97**	**1.00**	**0.96**
	#50	0.79	0.57	**0.91**	**0.97**	**1.00**	**0.96**
ReliefF	#10	**0.84**	**1.00**	0.75	0.94	0.88	**0.96**
	#50	**0.84**	0.85	0.83	**0.97**	**1.00**	**0.96**
SVM-RFE	#10	0.68	**1.00**	0.50	0.79	**1.00**	0.72
	#50	0.68	**1.00**	0.50	0.74	**1.00**	0.64
mRMR	#10	0.63	0.71	0.58	0.44	**1.00**	0.24
	#50	0.68	0.71	0.66	0.91	0.77	**0.96**

It is worth noting that the embedded method SVM-RFE, in spite of the fact that it is, in theory, better than the filters, achieves comparable or even worse results than them in terms of classification accuracy. Even when combined with the SVM classifier (see Table 12) it does not obtain the highest accuracy, contrary to what was expected. However, two datasets are not a sufficient basis from which to draw strong conclusions so it is necessary to check the results on the remaining datasets.

5.7. Cross-validation study

This section shows the classification results obtained when applying the well-known cross-validation technique. To this end, a 5-fold cross-validation was performed over the binary datasets presented in Table 8, which only have the training set available. Since in some cases cross-validation can potentially introduce the dataset shift problem, another strategy has been used. Distribution optimally balanced stratified cross-validation (DOB-SCV) tries to avoid dataset shift by assigning close-by examples to different folds.

Section 5.7.1 analyzes the behavior of the feature selection methods studied on the datasets, whilst Section 5.7.2 compares the performance of regular cross-validation against DOB-SCV. Finally, Section 5.7.3 will analyze the influence of the datasets' characteristics. Notice that due to the large amount of results obtained, the tables that will be analyzed in this section have been moved to Appendix A for the sake of clarity.

5.7.1. Analysis of algorithms

This subsection aims to analyze the behavior of the feature selection methods as well as the influence of the classifier on the studied datasets. Some interesting conclusions can be extracted by looking at the results reported by Tables A.1, A.2, A.3, A.4, A.5 and A.6.

1. The best performances are obtained by SVM and naive Bayes classifiers. As mentioned above, some studies [47] noted the superiority of SVMs over other classifiers. On the other hand, the performance of C4.5 may be affected by its embedded feature selection, in some cases leading to an extremely reduced set of features which can degrade the classification accuracy.

2. Focusing on the feature selection methods, on average for all datasets, the subset filters show an outstanding behavior, particularly CFS and INTERACT. It is surprising that SVM-RFE did not achieve the best results when combined with the SVM classifier, but the poor performance of the ranker methods can be explained by the restriction of having to establish a threshold for the number of features to retain. In the case of the subset filters, the number of features which form the final subset of features is the optimal one for a given dataset and method. However, the main disadvantage of rankers is the necessity of setting the threshold a priori. This has the risk of choosing too large or too small a number.
3. All the methods tested except Information Gain are multivariate, which in theory should show a better performance than univariate methods. However, on average, for all the datasets evaluated, Information Gain obtains similar classification accuracy results to the other algorithms, considering that its complexity is lower.
4. Some of the algorithms employed in this experimental study are based on information theory (Information Gain, mRMR, FCBF and INTERACT) whilst the rest are based on correlation coefficients (CFS, ReliefF and SVM-RFE). It seems that there is no relation between the nature of the methods and their behavior. In fact, CFS achieves the highest accuracy on average for most of the classifiers, whilst ReliefF obtains in many cases the poorest results, and both of them are based on correlation coefficients.

5.7.2. Cross-validation vs. DOB-SCV

The purpose of this subsection is to check the appropriateness of performing DOB-SCV instead of regular 5-fold cross-validation. On average, for all datasets, DOB-SCV obtains better results than those of regular cross-validation for SVM and naive Bayes classifiers, which are the classifiers which showed the best overall performance. It is interesting to focus on the case of the Brain dataset, which presents a high imbalance and an important amount of overlapping, as can be seen in Table 8. For this dataset, DOB-SCV outperforms regular cross-validation for SVM and naive Bayes classifiers, although the highest accuracy was achieved by C4.5 combined with the regular cross-validation. In the case of CNS, another complex dataset, there are no important differences between the two validation methods. On the other hand, there is the DLBCL dataset, which is in theory a simpler dataset (see Table 8). In fact, the accuracy obtained by the classifiers is, in most cases around 90%. Nevertheless, it is interesting to note that for this dataset, DOB-SCV outperforms, on average, the regular cross-validation. These datasets will be studied in detail in Section 5.7.3.

5.7.3. Analysis of datasets' characteristics

As mentioned in Section 4.3, microarray datasets present several problematics such as the imbalance of the data, the overlapping between classes or the dataset shift. Tables 7 and 8 reported the imbalance ratio and F1 of the datasets studied

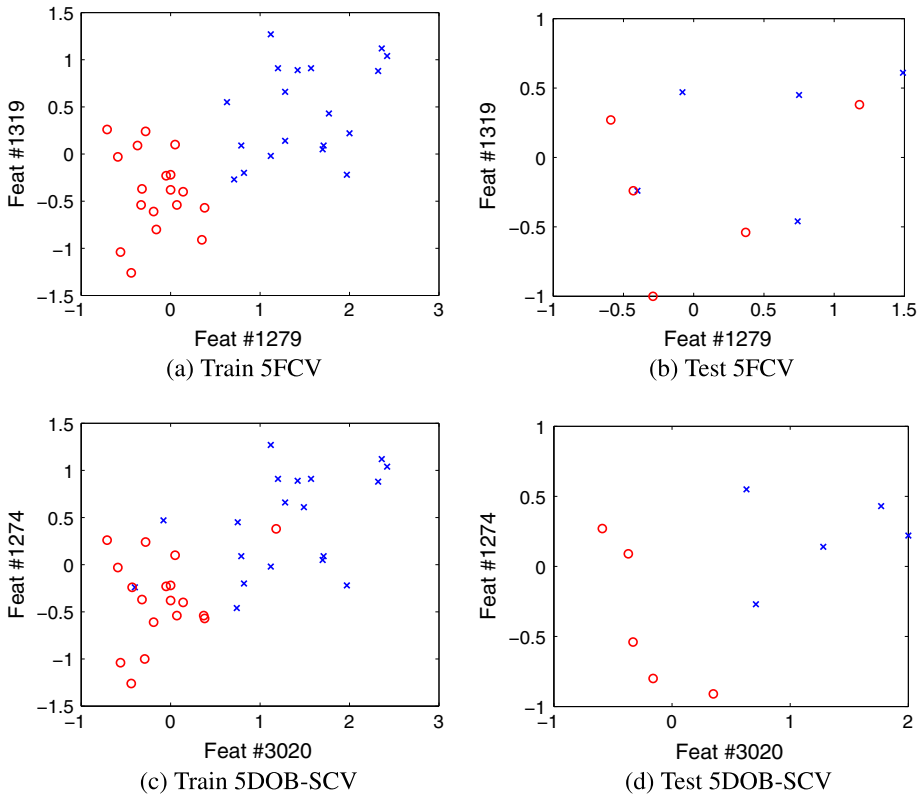

Fig. 4. Two first features selected by mRMR in the first fold for both 5-fold cross-validation and 5DOB-SCV.

in this section, which measures the imbalance of the data and the overlapping, respectively. Gli85 is the most unbalanced dataset, and its highest accuracy was obtained by SVM with a regular 5-fold cross validation and no feature selection (92%), although the Information Gain filter achieves 91% in accuracy and this degradation can be equivalent to the missclassification of only one or two samples.

SMK is the dataset which presents the highest level of overlapping between classes, and its maximum classification accuracy is very poor, around 70%. A similar case happens with the CNS dataset, which also has a low value of F1 (see Table 8).

Regarding the dataset shift problem and the adequacy of DOB-SCV, we will analyze in detail the case of the DLBCL dataset. Fig. 4 displays the 2-D representation of the first two features selected by mRMR in the first fold of a 5-fold cross-validation and a 5DOB-SCV for both train and test sets, in which different colors stand for different classes. As can be seen, cross-validation can indeed introduce dataset shift. The two first features selected by mRMR obtain a linearly separable problem (see Fig. 4(a)) in the train set, but these features are not so informative in the test set (see Fig. 4(b)). However, the partitions created by DOB-SCV do not suffer from dataset shift. In Fig. 4(c), the first two features selected by mRMR in the training set make the problem almost linearly separable and this condition is maintained in the test set. In fact, it has been demonstrated in the previous section that DOB-SCV outperformed the regular cross-validation for this dataset.

5.8. Summary of results

In light of the above, it can be seen that the results obtained by this experimental study are highly dependent on the classifier, the feature selection method, and in particular the dataset. Although a detailed analysis of the results is outside the scope of this paper, it is evident that the large number of problematics present in this type of dataset make the classification task very arduous. In such a situation, the authors recommend the careful study of the particularities of each problem, although it seems that the best results (in general) are obtained with the SVM classifier, a subset filter for feature selection, and the DOB-SCV validation method.

6. Conclusions

This article reviews the up-to-date contributions of feature selection research applied to the field of DNA microarray data analysis, as well as the datasets used. The advent of this type of data has posed a big challenge for machine learning researchers, because of the large input dimensionality and small sample size. Since the infancy of microarray data classification, feature selection has become an imperative step, in order to reduce the number of features (genes).

Since the end of the 1990s, when microarray datasets began to be dealt with, a large number of feature selection methods have been applied. In the literature, one can find both classical methods and methods developed especially for this kind of data. Due to the high computational resources that these datasets demand, wrapper and embedded methods have mostly been avoided, in favor of less expensive approaches such as filters.

The key point to understanding all the attention devoted to microarray data is the challenge that their problematic poses. Besides the obvious disadvantage of having so many features for such a small number of samples, researchers also have to deal with classes that are very unbalanced, training and test datasets extracted under different conditions, dataset shift or the presence of outliers. For these reasons, new methods emerge every year, not only in an attempt to improve previous results in terms of classification accuracy, but also aiming to help biologists identify the underlying mechanism that relates gene expression to diseases.

The objective of this paper is to review the characteristics of microarray datasets and the feature selection methods applied to this domain, gathering as much up-to-date knowledge as possible for the interested reader. Bearing this in mind, we have analyzed the recent literature in order to describe in broad brushstrokes the trends in the development of feature selection methods for microarray data. Furthermore, a summary of the datasets used in recent years is provided. In order to present a complete picture of the topic, we have also mentioned the most common validation techniques. Since there is no consensus in the literature about this issue, we have provided some guidelines. Other important aspects to consider when dealing with microarray data classification are model and classifier selection, parameter optimization or performance measures. However, this paper is focused on the difficulties derived from the data itself, thus the previously mentioned issues are outside the scope of this paper.

Finally, we have performed a practical evaluation for feature selection methods using microarray datasets in which we analyze the results obtained. This experimental study tries to show in practice the problematics that we have explained in theory. To this end, a suite of 9 widely-used binary datasets was chosen to apply over them 7 classical feature selection methods. In order to obtain the final classification accuracy, 3 well-known classifiers were used. This large set of experiments also aims at facilitating future comparative studies when a researcher proposes a new method.

Regarding the opportunities for future feature selection research, the tendency is toward focusing on new combinations such as hybrid or ensemble methods. These types of methods are able to enhance the robustness of the final subset of selected features, which is also a trending topic in this domain. Another interesting line of future research might be to distribute the microarray data vertically (i.e. by features) in order to reduce the heavy computational burden when applying wrapper methods.

Table A.1
Experimental results for C4.5 classifier on binary datasets after performing regular 5-fold cross-validation.

			Brain	CNS	Colon	DLBCL	Gli85	Ovarian	Smk	Avg
No FS		Ac	**1.00**	0.58	0.74	0.70	0.75	0.97	0.65	0.77
		Se	**1.00**	0.64	0.60	0.69	0.81	0.95	0.66	0.77
		Sp	**1.00**	0.48	0.82	0.70	0.63	0.98	0.62	0.75
CFS		Ac	**1.00**	0.62	0.79	0.75	0.79	0.98	0.64	**0.79**
		Se	**1.00**	0.64	0.68	0.78	0.81	0.95	0.56	0.78
		Sp	**1.00**	**0.58**	0.85	0.71	0.75	**0.99**	**0.71**	**0.80**
FCBF		Ac	0.86	0.48	0.79	0.73	0.82	**0.99**	0.61	0.75
		Se	0.80	0.49	0.64	0.74	0.86	**0.99**	0.65	0.74
		Sp	0.86	0.50	0.87	0.70	0.75	**0.99**	0.56	0.75
INT		Ac	**1.00**	0.55	0.79	0.70	0.78	0.98	0.59	0.77
		Se	**1.00**	0.54	0.72	0.74	0.81	0.98	0.51	0.76
		Sp	**1.00**	**0.58**	0.82	0.66	0.71	0.98	0.66	0.77
IG	#10	Ac	0.71	0.62	0.72	0.75	**0.85**	0.96	0.60	0.74
		Se	0.70	0.69	0.78	0.79	0.88	0.93	0.71	0.78
		Sp	0.70	0.48	0.70	0.71	**0.79**	0.97	0.48	0.69
	#50	Ac	0.81	0.63	**0.84**	0.73	0.81	0.96	0.65	0.78
		Se	0.70	0.67	**0.83**	0.69	0.86	0.96	0.62	0.76
		Sp	0.87	**0.58**	0.85	0.74	0.71	0.97	0.67	0.77
ReliefF	#10	Ac	0.72	0.47	0.72	**0.85**	**0.85**	0.97	0.65	0.75
		Se	0.20	0.59	0.50	0.83	0.88	0.94	**0.80**	0.68
		Sp	**1.00**	0.25	0.85	**0.87**	0.77	**0.99**	0.47	0.74
	#50	Ac	0.62	0.53	0.82	0.73	0.82	**0.99**	0.61	0.73
		Se	0.20	0.60	0.68	0.74	0.88	**0.99**	0.61	0.67
		Sp	0.86	0.44	**0.90**	0.70	0.70	**0.99**	0.62	0.74
SVM-RFE	#10	Ac	0.57	**0.65**	0.71	0.81	0.81	0.98	0.60	0.73
		Se	0.00	**0.74**	0.60	0.82	0.85	0.98	0.65	0.66
		Sp	0.87	0.48	0.77	0.79	0.75	0.98	0.55	0.74
	#50	Ac	0.70	0.57	0.80	0.82	0.79	0.98	0.65	0.76
		Se	**1.00**	0.61	0.77	**0.84**	0.83	**0.99**	0.62	**0.81**
		Sp	0.56	0.49	0.82	0.79	0.70	0.98	0.66	0.72
mRMR	#10	Ac	0.86	0.55	0.82	0.75	0.79	0.98	**0.68**	0.77
		Se	0.90	0.72	0.68	0.79	0.86	0.96	0.71	0.80
		Sp	0.87	0.23	**0.90**	0.70	0.61	**0.99**	0.64	0.70
	#50	Ac	0.86	0.58	0.82	0.73	0.80	0.97	0.62	0.77
		Se	0.90	0.70	0.77	0.69	**0.91**	0.96	0.66	0.80
		Sp	0.87	0.39	0.85	0.74	0.54	0.98	0.57	0.71

Table A.2
Experimental results for C4.5 classifier on binary datasets after performing DOB-SCV with 5 folds.

			Brain	CNS	Colon	DLBCL	Gli85	Ovarian	Smk	Avg
No FS		Ac	**0.92**	0.52	0.72	0.72	0.77	0.96	0.56	0.74
		Se	0.80	0.58	0.55	0.70	0.79	0.93	0.63	0.71
		Sp	**1.00**	0.39	0.80	0.74	0.74	0.97	0.48	0.73
CFS		Ac	**0.92**	0.52	0.79	0.80	0.75	0.96	0.59	0.76
		Se	0.80	0.53	0.65	**0.82**	0.77	0.93	0.59	0.73
		Sp	**1.00**	0.48	0.87	0.78	0.70	**0.98**	0.59	0.77
FCBF		Ac	0.72	0.52	0.81	0.72	0.76	**0.98**	0.59	0.73
		Se	0.70	0.56	0.69	0.70	0.79	**0.98**	0.53	0.71
		Sp	0.76	0.44	0.87	0.74	0.70	**0.98**	0.65	0.74
INT		Ac	**0.92**	0.53	0.81	0.74	0.71	0.97	0.67	0.76
		Se	0.80	0.51	0.69	0.65	0.69	0.94	0.58	0.69
		Sp	**1.00**	**0.56**	0.87	0.84	0.74	**0.98**	**0.75**	**0.82**
IG	#10	Ac	0.72	0.60	0.79	0.78	0.82	0.96	0.62	0.76
		Se	0.80	0.67	0.59	0.74	0.81	0.94	0.61	0.74
		Sp	0.73	0.49	**0.90**	0.83	**0.85**	0.96	0.63	0.77
	#50	Ac	0.77	0.54	**0.84**	0.80	0.81	**0.98**	0.60	0.76
		Se	0.80	0.59	**0.74**	**0.82**	0.79	**0.98**	0.58	0.76
		Sp	0.80	0.44	**0.90**	0.79	**0.85**	0.97	0.62	0.77

Table A.2 (*continued*)

			Brain	CNS	Colon	DLBCL	Gli85	Ovarian	Smk	Avg
ReliefF	#10	Ac	0.48	0.55	0.77	**0.84**	0.81	0.96	0.66	0.73
		Se	0.10	**0.76**	0.69	**0.82**	**0.91**	0.93	0.67	0.70
		Sp	0.63	0.15	0.82	0.87	0.57	0.98	0.64	0.67
	#50	Ac	0.53	0.56	0.76	0.80	**0.85**	**0.98**	0.68	0.74
		Se	0.60	0.69	0.63	**0.82**	0.86	**0.98**	**0.75**	0.76
		Sp	0.50	0.34	0.82	0.79	0.81	**0.98**	0.60	0.69
SVM-RFE	#10	Ac	0.59	0.56	0.76	0.82	0.78	**0.98**	0.60	0.73
		Se	0.40	0.66	0.64	0.73	0.79	0.96	0.66	0.69
		Sp	0.70	0.37	0.82	**0.92**	0.74	**0.98**	0.53	0.72
	#50	Ac	0.70	**0.62**	0.78	0.78	0.76	0.97	0.65	0.75
		Se	0.70	0.65	0.56	0.69	0.76	0.95	0.69	0.71
		Sp	0.73	**0.56**	**0.90**	0.88	0.73	**0.98**	0.61	0.77
mRMR	#10	Ac	0.80	0.60	0.79	0.78	0.80	**0.98**	0.68	**0.78**
		Se	0.80	0.71	0.65	0.78	0.79	**0.98**	**0.75**	**0.78**
		Sp	0.83	0.38	0.87	0.80	0.81	**0.98**	0.61	0.75
	#50	Ac	0.84	0.60	0.82	0.76	0.78	**0.98**	**0.71**	**0.78**
		Se	**0.90**	0.66	0.69	0.78	0.76	**0.98**	0.71	**0.78**
		Sp	0.83	0.47	**0.90**	0.74	0.82	**0.98**	0.70	0.78

Table A.3
Experimental results for naive Bayes classifier on binary datasets after performing regular 5-fold cross-validation.

			Brain	CNS	Colon	DLBCL	Gli85	Ovarian	Smk	Avg
No FS		Ac	0.67	0.60	0.55	0.92	0.84	0.93	0.63	0.73
		Se	0.00	0.64	0.69	**0.96**	0.88	0.99	0.60	0.68
		Sp	**1.00**	0.52	0.47	0.88	0.73	0.89	0.66	0.74
CFS		Ac	0.81	0.67	**0.85**	0.90	0.82	**1.00**	0.65	**0.81**
		Se	0.50	0.75	0.76	**0.96**	**0.90**	0.99	0.67	0.79
		Sp	**1.00**	0.54	**0.90**	0.84	0.67	**1.00**	0.62	0.79
FCBF		Ac	0.61	**0.70**	0.80	0.90	0.85	0.99	0.69	0.79
		Se	**1.00**	0.77	0.76	**0.96**	**0.90**	**1.00**	0.72	**0.87**
		Sp	0.40	0.58	0.82	0.84	0.74	0.99	0.65	0.72
INT		Ac	0.81	**0.70**	0.77	0.90	0.82	**1.00**	0.64	**0.81**
		Se	0.50	0.77	0.76	**0.96**	0.88	**1.00**	0.72	0.80
		Sp	**1.00**	0.58	0.77	0.83	0.71	0.99	0.55	0.78
IG	#10	Ac	**0.86**	0.63	0.79	**0.94**	0.85	0.96	0.61	**0.81**
		Se	0.70	0.67	0.72	**0.96**	0.88	0.95	0.59	0.78
		Sp	0.93	0.58	0.82	**0.92**	0.77	0.96	0.64	0.80
	#50	Ac	0.81	0.63	0.77	0.92	0.85	0.98	0.66	0.80
		Se	0.50	0.75	0.76	**0.96**	0.86	0.96	0.67	0.78
		Sp	**1.00**	0.42	0.77	0.88	0.81	0.98	0.65	0.79
ReliefF	#10	Ac	0.20	0.63	0.82	**0.94**	0.86	0.96	0.67	0.73
		Se	0.20	0.72	0.72	**0.96**	0.88	0.95	0.71	0.73
		Sp	0.20	0.48	0.87	**0.92**	0.81	0.96	0.63	0.70
	#50	Ac	0.19	0.67	0.84	0.92	**0.89**	0.98	0.67	0.74
		Se	0.50	0.72	0.77	**0.96**	0.86	0.95	0.72	0.78
		Sp	0.07	0.58	0.87	0.88	**0.97**	0.99	0.61	0.71
SVM-RFE	#10	Ac	0.62	0.68	0.73	0.92	0.82	0.99	**0.71**	0.78
		Se	0.30	0.77	0.61	0.91	0.83	**1.00**	**0.77**	0.74
		Sp	0.76	0.54	0.80	**0.92**	0.81	0.98	0.64	0.78
	#50	Ac	0.67	**0.70**	0.76	0.92	0.88	0.99	0.70	0.80
		Se	0.20	**0.82**	0.69	0.91	0.86	**1.00**	0.73	0.75
		Sp	0.90	0.49	0.80	**0.92**	0.93	0.98	0.65	**0.81**
mRMR	#10	Ac	0.73	0.63	0.80	0.92	0.85	0.99	0.67	0.80
		Se	0.60	0.79	0.78	**0.96**	0.88	0.96	0.68	0.81
		Sp	0.86	0.33	0.82	0.88	0.77	**1.00**	0.65	0.76
	#50	Ac	0.63	0.62	0.80	**0.94**	0.80	0.99	0.67	0.78
		Se	0.20	0.75	**0.86**	**0.96**	0.81	0.98	0.67	0.75
		Sp	0.86	0.38	0.77	**0.92**	0.77	0.99	**0.67**	0.77

Table A.4
Experimental results for naive Bayes classifier on binary datasets after performing DOB-SCV with 5 folds.

			Brain	CNS	Colon	DLBCL	Gli85	Ovarian	Smk	Avg
No FS		Ac	0.67	0.63	0.58	**0.98**	0.87	0.94	0.63	0.76
		Se	0.00	0.69	0.72	**1.00**	0.91	0.98	0.58	0.70
		Sp	**1.00**	0.54	0.50	0.96	0.77	0.91	0.67	0.77
CFS		Ac	**0.92**	0.67	0.82	0.93	0.85	0.98	0.68	**0.84**
		Se	**0.80**	0.76	0.77	**1.00**	**0.93**	0.98	0.74	0.85
		Sp	**1.00**	0.50	0.85	0.87	0.66	0.98	0.62	0.78
FCBF		Ac	0.71	0.58	0.79	0.96	**0.88**	0.99	0.66	0.80
		Se	**0.80**	0.69	0.77	**1.00**	**0.93**	0.99	0.70	0.84
		Sp	0.70	0.40	0.80	0.92	0.77	**0.99**	0.62	0.74
INT		Ac	**0.92**	0.62	**0.84**	0.93	0.86	0.99	0.70	**0.84**
		Se	**0.80**	0.69	**0.86**	**1.00**	0.91	**1.00**	0.74	**0.86**
		Sp	**1.00**	0.50	0.82	0.87	0.73	**0.99**	0.65	0.79
IG	#10	Ac	**0.92**	0.60	0.77	0.91	0.87	0.96	0.68	0.82
		Se	**0.80**	0.59	0.72	0.96	0.89	0.95	0.66	0.80
		Sp	**1.00**	**0.64**	0.80	0.87	0.81	0.96	**0.71**	**0.83**
	#50	Ac	**0.92**	0.65	0.79	0.96	0.87	0.98	0.70	**0.84**
		Se	**0.80**	0.72	0.82	0.96	0.90	0.96	0.68	0.83
		Sp	**1.00**	0.54	0.77	0.96	0.81	**0.99**	**0.71**	**0.83**
ReliefF	#10	Ac	0.26	0.65	**0.84**	0.93	0.84	0.96	0.67	0.74
		Se	0.30	0.69	0.72	0.96	0.88	0.95	0.71	0.74
		Sp	0.20	0.57	**0.90**	0.91	0.74	0.96	0.62	0.70
	#50	Ac	0.21	0.67	**0.84**	0.96	0.86	0.98	0.68	0.74
		Se	0.30	0.71	0.72	0.96	0.86	0.95	**0.77**	0.75
		Sp	0.13	0.58	**0.90**	0.96	**0.85**	**0.99**	0.59	0.71
SVM-RFE	#10	Ac	0.58	**0.72**	0.76	0.94	0.83	**1.00**	0.70	0.79
		Se	0.20	0.76	0.64	0.91	0.86	**1.00**	0.75	0.73
		Sp	0.73	0.63	0.82	0.96	0.74	**0.99**	0.63	0.79
	#50	Ac	0.71	0.69	0.76	0.92	0.87	0.99	0.68	0.80
		Se	0.10	**0.77**	0.69	**1.00**	0.88	**1.00**	0.76	0.74
		Sp	**1.00**	0.54	0.80	0.84	**0.85**	**0.99**	0.59	0.80
mRMR	#10	Ac	0.75	0.68	0.82	**0.98**	0.87	0.99	**0.71**	0.83
		Se	**0.80**	0.74	0.77	**1.00**	**0.93**	0.98	0.74	0.85
		Sp	0.73	0.58	0.85	0.96	0.73	**0.99**	0.66	0.79
	#50	Ac	0.77	0.67	0.79	**0.98**	0.85	0.98	0.67	0.82
		Se	0.40	0.71	0.82	0.96	0.87	0.96	0.70	0.78
		Sp	**1.00**	0.59	0.77	**1.00**	0.77	**0.99**	0.64	0.82

Table A.5
Experimental results for SVM classifier on binary datasets after performing regular 5-fold cross-validation.

			Brain	CNS	Colon	DLBCL	Gli85	Ovarian	Smk	Avg
No FS		Ac	**0.68**	0.67	0.77	**0.96**	**0.92**	**1.00**	**0.72**	**0.82**
		Se	0.20	0.82	0.60	0.96	**0.98**	**1.00**	0.79	0.77
		Sp	0.93	0.38	0.87	**0.96**	0.78	**1.00**	0.63	**0.79**
CFS		Ac	0.61	0.62	0.81	0.88	0.88	**1.00**	0.64	0.78
		Se	0.60	0.70	0.69	0.86	0.93	**1.00**	0.66	0.78
		Sp	0.66	0.49	0.87	0.88	0.77	**1.00**	0.61	0.76
FCBF		Ac	0.67	0.65	0.84	0.81	0.87	**1.00**	0.71	0.79
		Se	0.00	0.80	0.73	0.82	0.93	**1.00**	0.76	0.72
		Sp	**1.00**	0.39	0.90	0.79	0.73	**1.00**	0.64	0.78
INT		Ac	0.61	0.62	0.81	0.88	0.88	**1.00**	0.66	0.78
		Se	0.60	0.75	0.64	0.91	0.91	**1.00**	0.69	0.79
		Sp	0.66	0.39	0.90	0.83	0.81	**1.00**	0.63	0.75
IG	#10	Ac	0.48	0.63	0.81	0.94	0.91	0.98	0.64	0.77
		Se	0.00	0.82	0.59	0.96	**0.98**	0.96	0.74	0.72
		Sp	0.70	0.30	**0.92**	0.92	0.74	0.99	0.53	0.73
	#50	Ac	0.66	0.67	**0.85**	0.94	0.86	**1.00**	**0.72**	0.81
		Se	**0.80**	0.77	**0.81**	0.96	0.90	**1.00**	0.73	**0.85**
		Sp	0.66	0.48	0.87	0.92	0.77	0.99	0.70	0.77
ReliefF	#10	Ac	0.50	0.68	0.81	0.94	0.87	0.98	0.69	0.78
		Se	0.00	0.87	0.60	0.96	0.96	0.94	**0.82**	0.74

Table A.5 (*continued*)

			Brain	CNS	Colon	DLBCL	Gli85	Ovarian	Smk	Avg
		Sp	0.73	0.34	**0.92**	0.92	0.66	0.99	0.54	0.73
	#50	Ac	0.35	**0.73**	**0.85**	0.92	0.89	**1.00**	0.69	0.78
		Se	0.00	0.82	0.72	**1.00**	0.93	**1.00**	0.74	0.74
		Sp	0.53	**0.58**	0.92	0.84	0.82	**1.00**	0.64	0.76
SVM-RFE	#10	Ac	0.62	**0.73**	0.73	0.87	0.86	**1.00**	0.70	0.79
		Se	0.20	0.84	0.56	0.87	0.88	**1.00**	0.78	0.73
		Sp	0.86	0.53	0.82	0.88	0.81	**1.00**	0.61	**0.79**
	#50	Ac	0.48	0.72	0.71	0.88	0.89	**1.00**	**0.72**	0.77
		Se	0.20	0.82	0.57	0.91	0.91	**1.00**	0.74	0.74
		Sp	0.63	0.53	0.80	0.84	**0.85**	**1.00**	**0.68**	0.76
mRMR	#10	Ac	0.53	0.65	0.77	0.92	0.89	**1.00**	0.68	0.78
		Se	0.60	**0.95**	0.56	0.96	0.95	0.99	0.74	0.82
		Sp	0.56	0.10	0.90	0.87	0.77	**1.00**	0.62	0.69
	#50	Ac	0.49	0.70	0.84	**0.96**	0.89	**1.00**	0.68	0.79
		Se	0.20	0.77	0.77	**1.00**	0.95	**1.00**	0.74	0.78
		Sp	0.63	0.57	0.87	0.92	0.77	**1.00**	0.62	0.77

Table A.6
Experimental results for SVM classifier on binary datasets after performing DOB-SCV with 5 folds.

			Brain	CNS	Colon	DLBCL	Gli85	Ovarian	Smk	Avg
No FS		Ac	0.67	0.65	0.84	0.93	**0.91**	**1.00**	**0.72**	0.82
		Se	0.30	0.77	**0.82**	0.95	0.97	**1.00**	0.77	0.80
		Sp	0.86	0.44	0.85	0.92	0.77	**1.00**	0.66	**0.79**
CFS		Ac	**0.80**	0.66	0.82	0.94	0.89	**1.00**	0.68	**0.83**
		Se	0.70	0.74	0.78	**1.00**	0.97	0.99	0.70	0.84
		Sp	0.87	0.54	0.85	0.88	0.73	**1.00**	0.66	**0.79**
FCBF		Ac	0.67	0.58	0.79	0.92	0.90	0.99	0.66	0.79
		Se	0.00	0.71	0.68	**1.00**	0.93	0.98	0.68	0.71
		Sp	**1.00**	0.35	0.85	0.84	**0.81**	**1.00**	0.63	0.78
INT		Ac	**0.80**	0.60	0.77	0.91	0.87	**1.00**	**0.72**	0.81
		Se	0.70	0.66	0.63	0.95	0.91	0.99	0.74	0.80
		Sp	0.87	0.49	0.85	0.88	0.77	**1.00**	0.68	**0.79**
IG	#10	Ac	0.62	0.65	0.79	0.91	**0.91**	0.96	0.68	0.79
		Se	0.00	0.81	0.58	0.96	**0.98**	0.94	0.67	0.71
		Sp	0.93	0.34	0.90	0.87	0.73	0.97	0.70	0.78
	#50	Ac	0.66	0.67	**0.85**	**0.98**	0.87	**1.00**	0.65	0.81
		Se	**0.80**	0.74	**0.82**	**1.00**	0.93	**1.00**	0.70	**0.86**
		Sp	0.63	0.54	0.87	**0.96**	0.73	**1.00**	0.60	0.76
ReliefF	#10	Ac	0.45	0.63	0.82	0.95	0.84	0.98	0.65	0.76
		Se	0.10	0.84	0.58	**1.00**	0.93	0.93	0.75	0.73
		Sp	0.60	0.25	**0.95**	0.91	0.62	**1.00**	0.54	0.70
	#50	Ac	0.64	0.63	0.84	0.94	0.88	0.99	**0.72**	0.81
		Se	0.30	0.74	0.81	**1.00**	0.95	0.98	**0.83**	0.80
		Sp	0.80	0.43	0.85	0.88	0.73	**1.00**	0.58	0.75
SVM-RFE	#10	Ac	0.62	0.67	0.76	0.94	0.85	**1.00**	0.67	0.79
		Se	0.10	0.71	0.60	**1.00**	0.95	**1.00**	0.74	0.73
		Sp	0.86	0.58	0.85	0.88	0.61	**1.00**	0.60	0.77
	#50	Ac	0.67	0.65	0.81	0.91	0.86	**1.00**	0.71	0.80
		Se	0.30	0.72	0.78	0.95	0.91	**1.00**	0.70	0.77
		Sp	0.86	0.54	0.82	0.88	0.73	**1.00**	**0.71**	**0.79**
mRMR	#10	Ac	0.45	0.65	0.82	**0.98**	0.87	**1.00**	0.71	0.78
		Se	0.20	**0.87**	0.72	**1.00**	0.95	**1.00**	0.74	0.78
		Sp	0.60	0.24	0.87	**0.96**	0.69	**1.00**	0.66	0.72
	#50	Ac	0.58	**0.70**	0.84	0.93	0.87	**1.00**	0.66	0.80
		Se	0.10	0.74	0.78	0.95	0.95	**1.00**	0.73	0.75
		Sp	0.80	**0.63**	0.87	0.92	0.70	**1.00**	0.57	0.78

Acknowledgments

This research has been economically supported in part by the Secretaría de Estado de Investigación of the Spanish Government through the research Projects TIN 2011-28488 and TIN 2012-37954; by the Consellería de Industria of the Xunta de Galicia through the research Projects CN2011/007 and CN2012/211; and by the regional Project P11-TIC-9704; all of them partially funded by FEDER funds of the European Union. V. Bolón-Canedo acknowledges the support of Xunta de Galicia under *Plan I2C* Grant Program and the Genil Grant, which provided a stay in the Granada Excellence Network of Innovation Laboratories.

Appendix A. Tables for cross-validation study

This appendix reports the experimental results achieved for the cross-validation study. Tables A.1–A.6 are devoted to the three classifiers employed (C4.5, naive Bayes and SVM) and the two types of cross validation evaluated: the regular one and DOB-SCV. In both cases, 5 folds were considered and the results shown in the tables are the average test results for the 5 folds. These tables depict the classification accuracy (Ac), sensitivity (Se) and specificity (Sp). For the sake of comparison, the first row reports those values without applying feature selection. The best results for dataset and classifier regarding each measure are marked in bold face.

References

[1] Arrayexpress – Functional Genomics Data. <http://www.ebi.ac.uk/arrayexpress/> (accessed January, 2014).
[2] Broad institute. Cancer Program Data Sets. <http://www.broadinstitute.org/cgi-bin/cancer/datasets.cgi> (accessed January, 2014).
[3] Dataset Repository, Bioinformatics Research Group. <http://www.upo.es/eps/bigs/datasets.html> (accessed January, 2014).
[4] Feature Selection Algorithms at Arizona State University. <http://featureselection.asu.edu/software.php> (accessed January, 2014).
[5] Feature Selection Datasets at Arizona State University. <http://featureselection.asu.edu/datasets.php> (accessed January, 2014).
[6] Gene Expression Omnibus. <http://www.ncbi.nlm.nih.gov/geo/> (accessed January, 2014).
[7] Gene Expression Project, Princeton University. <http://genomics-pubs.princeton.edu/oncology/> (accessed January, 2014).
[8] Kent Ridge Bio-Medical Dataset. <http://datam.i2r.a-star.edu.sg/datasets/krbd> (accessed January, 2014).
[9] Microarray Cancers, Plymouth University. <http://www.tech.plym.ac.uk/spmc/links/bioinformatics/microarray/microarray_cancers.html> (accessed January, 2014).
[10] Stanford Microarray Database. <http://smd.stanford.edu/> (accessed January, 2014).
[11] T. Abeel, T. Helleputte, Y. Van de Peer, P. Dupont, Y. Saeys, Robust biomarker identification for cancer diagnosis with ensemble feature selection methods, Bioinformatics 26 (3) (2000) 392–398.
[12] J. Alcalá-Fdez, L. Sánchez, S. García, M.J. del Jesús, S. Ventura, J. Garrell, J. Otero, C. Romero, J. Bacardit, V.M. Rivas, et al, Keel: a software tool to assess evolutionary algorithms for data mining problems, Soft Comput. 13 (3) (2009) 307–318.
[13] A. Alizadeh, M. Eisen, R. Davis, C. Ma, I. Lossos, A. Rosenwald, J. Boldrick, H. Sabet, T. Tran, X. Yu, et al, Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling, Nature 403 (6769) (2000) 503–511.
[14] U. Alon, N. Barkai, D. Notterman, K. Gish, S. Ybarra, D. Mack, A. Levine, Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, Proc. Nat. Acad. Sci. 96 (12) (1999) 6745–6750.
[15] C. Ambroise, G. McLachlan, Selection bias in gene extraction on the basis of microarray gene-expression data, Proc. Nat. Acad. Sci. 99 (10) (2002) 6562–6566.
[16] A. Anaissi, P.J. Kennedy, M. Goyal, Feature selection of imbalanced gene expression microarray data, in: Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), 2011 12th ACIS International Conference on, IEEE, 2011, pp. 73–78.
[17] S. Armstrong, J. Staunton, L. Silverman, R. Pieters, M. den Boer, M. Minden, S. Sallan, E. Lander, T. Golub, S. Korsmeyer, et al, Mll translocations specify a distinct gene expression profile that distinguishes a unique leukemia, Nat. Genet. 30 (1) (2002) 41–47.
[18] V. Barnett, T. Lewis, Outliers in Statistical Data, vol. 3, Wiley, New York, 1994.
[19] M.R. Berthold, N. Cebron, F. Dill, T.R. Gabriel, T. Kötter, T. Meinl, P. Ohl, C. Sieb, K. Thiel, B. Wiswedel, KNIME: The Konstanz Information Miner, in: Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007), Springer, 2007.
[20] A. Bhattacharjee, W. Richards, J. Staunton, C. Li, S. Monti, P. Vasa, C. Ladd, J. Beheshti, R. Bueno, M. Gillette, et al, Classification of human lung carcinomas by mrna expression profiling reveals distinct adenocarcinoma subclasses, Proc. Nat. Acad. Sci. 98 (24) (2001) 13790–13795.
[21] R. Blagus, L. Lusa, Evaluation of smote for high-dimensional class-imbalanced microarray data, 2012 11th International Conference on Machine Learning and Applications (ICMLA), vol. 2, IEEE, 2012, pp. 89–94.
[22] R. Blanco, P. Larrañaga, I. Inza, B. Sierra, Gene selection for cancer classification using wrapper approaches, Int. J. Pattern Recognit. Artif. Intell. 18 (08) (2004) 1373–1390.
[23] V. Bolón-Canedo, N. Sánchez-Maroño, A. Alonso-Betanzos, On the effectiveness of discretization on gene selection of microarray data, in: The 2010 International Joint Conference on Neural Networks (IJCNN), IEEE, 2010, pp. 18–23.
[24] V. Bolón-Canedo, N. Sánchez-Maroño, A. Alonso-Betanzos, An ensemble of filters and classifiers for microarray data classification, Pattern Recognit. 45 (1) (2012) 531–539.
[25] V. Bolón-Canedo, N. Sánchez-Maroño, A. Alonso-Betanzos, Data classification using an ensemble of filters, Neurocomputing 135 (2014) 13–20.
[26] V. Bolón-Canedo, N. Sánchez-Maroño, A. Alonso-Betanzos, A review of feature selection methods on synthetic data, Knowl. Inf. Syst. 34 (3) (2013) 483–519.
[27] V. Bolón-Canedo, S. Seth, A. Sánchez-Maroño, N. Alonso-Betanzos, J. Principe, Statistical dependence measure for feature selection in microarray datasets, in: 19th European Symposium on Artificial Neural Networks-ESANN, 2011, pp. 23–28.
[28] U. Braga-Neto, Fads and fallacies in the name of small-sample microarray classification-a highlight of misunderstanding and erroneous usage in the applications of genomic signal processing, IEEE Signal Process. Mag. 24 (1) (2007) 91–99.
[29] U. Braga-Neto, E. Dougherty, Is cross-validation valid for small-sample microarray classification?, Bioinformatics 20 (3) (2004) 374–380
[30] M. Brown, W. Grundy, D. Lin, N. Cristianini, C. Sugnet, T. Furey, M. Ares, D. Haussler, Knowledge-based analysis of microarray gene expression data by using support vector machines, Proc. Nat. Acad. Sci. 97 (1) (2000) 262–267.
[31] J. Canul-Reich, L. Hall, D. Goldgof, J. Korecki, S. Eschrich, Iterative feature perturbation as a gene selector for microarray data, Int. J. Pattern Recognit. Artif. Intell. 26 (05) (2012) 1260003.
[32] N. Chawla, K. Bowyer, L. Hall, W. Kegelmeyer, Smote: synthetic minority over-sampling technique, J. Artif. Intell. Res. 16 (2002) 321–357.
[33] L. Chuang, C. Yang, K. Wu, C. Yang, A hybrid feature selection method for dna microarray data, Comput. Biol. Med. 41 (4) (2011) 228–237.

[34] C. Ding, H. Peng, Minimum redundancy feature selection from microarray gene expression data, J. Bioinformatics Comput. Biol. 3 (02) (2005) 185–205.
[35] E. Dougherty, Small sample issues for microarray-based classification, Comp. Funct. Genom. 2 (1) (2001) 28–34.
[36] S. Dudoit, J. Fridlyand, T. Speed, Comparison of discrimination methods for the classification of tumors using gene expression data, J. Am. Stat. Assoc. 97 (457) (2002) 77–87.
[37] B. Efron, Bootstrap methods: another look at the jackknife, Ann. Stat. (1979) 1–26.
[38] S. Eschrich, I. Yang, G. Bloom, K. Kwong, D. Boulware, A. Cantor, D. Coppola, M. Kruhøffer, L. Aaltonen, T. Orntoft, et al, Molecular staging for survival prediction of colorectal cancer patients, J. Clin. Oncol. 23 (15) (2005) 3526–3535.
[39] A. Ferreira, M. Figueiredo, An unsupervised approach to feature discretization and selection, Pattern Recognit. 45 (9) (2012) 3048–3060.
[40] W. Freije, F. Castro-Vargas, Z. Fang, S. Horvath, T. Cloughesy, L. Liau, P. Mischel, S. Nelson, Gene expression profiling of gliomas strongly predicts survival, Cancer Res. 64 (18) (2004) 6503–6510.
[41] M. Galar, A. Fernández, E. Barrenechea, H. Bustince, F. Herrera, A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches, IEEE Trans. Syst. Man Cybern. Part C: Appl. Rev. 42 (4) (2012) 463–484.
[42] M. Galar, A. Fernández, E. Barrenechea, F. Herrera, Eusboost: Enhancing ensembles for highly imbalanced data-sets by evolutionary undersampling, Pattern Recognit. 46 (12) (2013) 3460–3471.
[43] M. Garber, O. Troyanskaya, K. Schluens, S. Petersen, Z. Thaesler, M. Pacyna-Gengelbach, M. Van De Rijn, G. Rosen, C. Perou, R. Whyte, et al, Diversity of gene expression in adenocarcinoma of the lung, Proc. Nat. Acad. Sci. 98 (24) (2001) 13784–13789.
[44] S. Garcıa, J. Luengo, J.A. Sáez, V. López, F. Herrera, A survey of discretization techniques: taxonomy and empirical analysis in supervised learning, IEEE Trans. Knowl. Data Eng. 25 (4) (2013).
[45] O. Gevaert, F. Smet, D. Timmerman, Y. Moreau, B. Moor, Predicting the prognosis of breast cancer by integrating clinical and microarray data with bayesian networks, Bioinformatics 22 (14) (2006) 184–190.
[46] T. Golub, D. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. Mesirov, H. Coller, M. Loh, J. Downing, M. Caligiuri, et al, Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, Science 286 (5439) (1999) 531–537.
[47] F. Gonzalez-Navarro, Feature Selection in Cancer Research: Microarray Gene Expression and in vivo 1H-MRS Domains. PhD thesis, Technical University of Catalonia, 2011.
[48] F. González Navarro, L. Belanche Muñoz, Gene subset selection in microarray data using entropic filtering for cancer classification, Expert Syst. 26 (1) (2009) 113–124.
[49] G. Gordon, R. Jensen, L. Hsiao, S. Gullans, J. Blumenstock, S. Ramaswamy, W. Richards, D. Sugarbaker, R. Bueno, Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma, Cancer Res. 62 (17) (2002) 4963–4967.
[50] I. Guyon, S. Gunn, M. Nikravesh, L. Zadeh, Feature Extraction: Foundations and Applications, vol. 207, Springer, 2006.
[51] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene selection for cancer classification using support vector machines, Mach. Learn. 46 (1-3) (2002) 389–422.
[52] M. Hall. Correlation-Based Feature Selection for Machine Learning. PhD thesis, Citeseer, 1999.
[53] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. Witten, The weka data mining software: an update, ACM SIGKDD Explor. Newsl. 11 (1) (2009) 10–18.
[54] M. Hall, L. Smith, Practical feature subset selection for machine learning, Comput. Sci. 98 (1998) 181–191.
[55] C. Haslinger, N. Schweifer, S. Stilgenbauer, H. Döhner, P. Lichter, N. Kraut, C. Stratowa, R. Abseher, Microarray gene expression profiling of b-cell chronic lymphocytic leukemia subgroups defined by genomic aberrations and vh mutation status, J. Clin. Oncol. 22 (19) (2004) 3937–3949.
[56] H. He, E.A. Garcia, Learning from imbalanced data, IEEE Trans. Knowl. Data Eng. 21 (9) (2009) 1263–1284.
[57] G. Heap, G. Trynka, R. Jansen, M. Bruinenberg, M. Swertz, L. Dinesen, K. Hunt, C. Wijmenga, et al, Complex nature of snp genotype effects on gene expression in primary human leucocytes, BMC Med. Genom. 2 (1) (2009) 1.
[58] I. Hedenfalk, D. Duggan, Y. Chen, M. Radmacher, M. Bittner, R. Simon, P. Meltzer, B. Gusterson, M. Esteller, M. Raffeld, et al, Gene-expression profiles in hereditary breast cancer, New England J. Med. 344 (8) (2001) 539–548.
[59] T.K. Ho, M. Basu, Complexity measures of supervised classification problems, IEEE Trans. Pattern Anal. Mach. Intell. 24 (3) (2002) 289–300.
[60] I. Inza, P. Larrañaga, R. Blanco, A. Cerrolaza, Filter versus wrapper gene selection approaches in dna microarray domains, Artif. Intell. Med. 31 (2) (2004) 91–103.
[61] I. Inza, B. Sierra, R. Blanco, P. Larrañaga, Gene selection by sequential search wrapper approaches in microarray cancer class prediction, J. Intell. Fuzzy Syst. 12 (1) (2002) 25–33.
[62] A. Jain, D. Zongker, Feature selection: evaluation, application, and small sample performance, IEEE Trans. Pattern Anal. Mach. Intell. 19 (2) (1997) 153–158.
[63] T. Jirapech-Umpai, S. Aitken, Feature selection and classification for microarray data analysis: evolutionary methods for identifying predictive genes, BMC Bioinformatics 6 (1) (2005) 148.
[64] N. Jojic, A. Perina, Multidimensional Counting Grids: Inferring Word Order from Disordered Bags of Words, 2012. arXiv preprint <1202.3752>.
[65] K. Kadota, D. Tominaga, Y. Akiyama, K. Takahashi, Detecting outlying samples in microarray data: a critical assessment of the effect of outliers on sample classification, Chem-Bio Infor. 3 (1) (2003) 30–45.
[66] J. Khan, J. Wei, M. Ringner, L. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. Antonescu, C. Peterson, et al, Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks, Nat. Med. 7 (6) (2001) 673–679.
[67] K. Kira, L. Rendell, The feature selection problem: traditional methods and a new algorithm, in: Proceedings of the National Conference on Artificial Intelligence, John Wiley & Sons Ltd, 1992. p. 129.
[68] I. Kononenko, Estimating attributes: analysis and extensions of relief, in: Machine Learning: ECML-94, Springer, 1994, pp. 171–182.
[69] L. Lan, S. Vucetic, Improving accuracy of microarray classification by a simple multi-task feature selection filter, Int. J. Data Mining Bioinformatics 5 (2) (2011) 189–208.
[70] C. Lazar, J. Taminau, S. Meganck, D. Steenhoff, A. Coletta, C. Molter, V. de Schaetzen, R. Duque, H. Bersini, A. Nowé, A survey on filter techniques for feature selection in gene expression microarray analysis, IEEE/ACM Trans. Comput. Biol. Bioinformatics (TCBB) 9 (4) (2012) 1106–1119.
[71] C. Lee, Y. Leu, A novel hybrid feature selection method for microarray data analysis, Appl. Soft Comput. 11 (1) (2011) 208–213.
[72] J. Lee, J. Lee, M. Park, S. Song, An extensive comparison of recent classification tools applied to microarray data, Comput. Statist. Data Anal. 48 (4) (2005) 869–885.
[73] Y. Leung, Y. Hung, A multiple-filter-multiple-wrapper approach to gene selection and microarray data classification, IEEE/ACM Trans. Comput. Biol. Bioinformatics (TCBB) 7 (1) (2010) 108–117.
[74] T. Li, C. Zhang, M. Ogihara, A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression, Bioinformatics 20 (15) (2004) 2429–2437.
[75] D. Liu, H. Qian, G. Dai, Z. Zhang, An iterative svm approach to feature selection and classification in high-dimensional datasets, Pattern Recognit. (2013).
[76] V. López, A. Fernández, S. García, V. Palade, F. Herrera, An insight into classification with imbalanced data: empirical results and current trends on using data intrinsic characteristics, Inf. Sci. 250 (0) (2013) 113–141.
[77] A.C. Lorena, I.G. Costa, N. Spolaôr, M.C. de Souto, Analysis of complexity indices for classification problems: cancer gene expression data, Neurocomputing 75 (1) (2012) 33–42.
[78] P. Lovato, M. Bicego, M. Cristani, N. Jojic, A. Perina, Feature selection using counting grids: application to microarray data, in: Structural, Syntactic, and Statistical Pattern Recognition, Springer, 2012, pp. 629–637.

[79] S. Ma, J. Huang, Penalized feature selection and classification in bioinformatics, Briefings Bioinformatics 9 (5) (2008) 392–403.
[80] S. Maldonado, R. Weber, J. Basak, Simultaneous feature selection and classification using kernel-penalized support vector machines, Inf. Sci. 181 (1) (2011) 115–128.
[81] P. Meyer, C. Schretter, G. Bontempi, Information-theoretic feature selection in microarray data using variable complementarity, IEEE J. Sel. Topics Signal Process. 2 (3) (2008) 261–274.
[82] S. Michiels, S. Koscielny, C. Hill, Prediction of cancer outcome with microarrays: a multiple random validation strategy, Lancet 365 (9458) (2005) 488–492.
[83] J.G. Moreno-Torres, T. Raeder, R. Alaiz-Rodríguez, N.V. Chawla, F. Herrera, A unifying view on dataset shift in classification, Pattern Recognit. 45 (1) (2012) 521–530.
[84] J.G. Moreno-Torres, J.A. Sáez, F. Herrera, Study on the impact of partition-induced dataset shift on k-fold cross-validation, IEEE Trans. Neural Networks Learn. Syst. 23 (8) (2012) 1304–1312.
[85] P. Mundra, J. Rajapakse, SVM-RFE with mRMR filter for gene selection, IEEE Trans. NanoBiosci. 9 (1) (2010) 31–37.
[86] F. Nie, H. Huang, X. Cai, C. Ding, Efficient and robust feature selection via joint l2, 1-norms minimization, Adv. Neural Inf. Process. Syst. 23 (2010) 1813–1821.
[87] C. Noble, A. Abbas, J. Cornelius, C. Lees, G. Ho, K. Toy, Z. Modrusan, N. Pal, F. Zhong, S. Chalasani, et al, Regional variation in gene expression in the healthy colon is dysregulated in ulcerative colitis, Gut 57 (10) (2008) 1398–1405.
[88] C. Nutt, D. Mani, R. Betensky, P. Tamayo, J. Cairncross, C. Ladd, U. Pohl, C. Hartmann, M.E. McLaughlin, T. Batchelor, et al, Gene expression-based classification of malignant gliomas correlates better with survival than histological classification, Cancer Res. 63 (7) (2003) 1602–1607.
[89] O. Okun, H. Priisalu, Dataset complexity in gene expression based cancer classification using ensembles of k-nearest neighbors, Artif. Intell. Med. 45 (2) (2009) 151–162.
[90] A. Orriols-Puig, E. Bernadó-Mansilla, Evolutionary rule-based systems for imbalanced data sets, Soft Comput. 13 (3) (2009) 213–225.
[91] H. Peng, F. Long, C. Ding, Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy, IEEE Trans. Pattern Anal. Mach. Intell. 27 (8) (2005) 1226–1238.
[92] T. Peters, D. Bulger, T. Loi, J. Yang, D. Ma, Two-step cross-entropy feature selection for microarrayspower through complementarity, IEEE/ACM Trans. Comput. Biol. Bioinformatics 8 (4) (2011) 1148–1151.
[93] E. Petricoin, A. Ardekani, B. Hitt, P. Levine, V. Fusaro, S. Steinberg, G. Mills, C. Simone, D. Fishman, E. Kohn, et al, Use of proteomic patterns in serum to identify ovarian cancer, Lancet 359 (9306) (2002) 572–577.
[94] G. Piatetsky-Shapiro, P. Tamayo, Microarray data mining: facing the challenges, ACM SIGKDD Explor. Newsl. 5 (2) (2003) 1–5.
[95] S. Pomeroy, P. Tamayo, M. Gaasenbeek, L. Sturla, M. Angelo, M. McLaughlin, J. Kim, L. Goumnerova, P. Black, C. Lau, et al, Prediction of central nervous system embryonal tumour outcome based on gene expression, Nature 415 (6870) (2002) 436–442.
[96] S. Ramaswamy, P. Tamayo, R. Rifkin, S. Mukherjee, C. Yeang, M. Angelo, C. Ladd, M. Reich, E. Latulippe, J. Mesirov, et al, Multiclass cancer diagnosis using tumor gene expression signatures, Proc. Nat. Acad. Sci. 98 (26) (2001) 15149–15154.
[97] R. Ruiz, J. Riquelme, J. Aguilar-Ruiz, Incremental wrapper-based gene selection from microarray data for cancer classification, Pattern Recognit. 39 (12) (2006) 2383–2392.
[98] Y. Saeys, I. Inza, P. Larrañaga, A review of feature selection techniques in bioinformatics, Bioinformatics 23 (19) (2007) 2507–2517.
[99] J.A. Saez, J. Luengo, F. Herrera, Predicting noise filtering efficacy with data complexity measures for nearest neighbor classification, Pattern Recognit. 46 (2013) 355–364.
[100] N. Sanchez-Marono, A. Alonso-Betanzos, P. Garcia-Gonzalez, V. Bolon-Canedo, Multiclass classifiers vs multiple binary classifiers using filters for feature selection, in: The 2010 International Joint Conference on Neural Networks (IJCNN), IEEE, 2010, pp. 18–23.
[101] M. Shah, M. Marchand, J. Corbeil, Feature selection with conjunctions of decision stumps and learning from microarray data, IEEE Trans. Pattern Anal. Mach. Intell. 34 (1) (2012) 174–186.
[102] A. Sharma, S. Imoto, S. Miyano, A top-r feature selection algorithm for microarray gene expression data, IEEE/ACM Trans. Comput. Biol. Bioinformatics (TCBB) 9 (3) (2012) 754–764.
[103] K. Shedden, J. Taylor, S. Enkemann, M. Tsao, T. Yeatman, W. Gerald, S. Eschrich, I. Jurisica, T. Giordano, D. Misek, et al, Gene expression–based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study, Nat. Med. 14 (8) (2008) 822–827.
[104] M. Shipp, K. Ross, P. Tamayo, A. Weng, J. Kutok, R. Aguiar, M. Gaasenbeek, M. Angelo, M. Reich, G. Pinkus, et al, Diffuse large b-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning, Nat. Med. 8 (1) (2002) 68–74.
[105] S. Shreem, S. Abdullah, M. Nazri, M. Alzaqebah, Hybridizing ReliefF, MRMR filters and GA wrapper approaches for gene selection, J. Theor. Appl. Inf. Technol. 46 (2) (2012) 1034–1039.
[106] D. Singh, P. Febbo, K. Ross, D. Jackson, J. Manola, C. Ladd, P. Tamayo, A. Renshaw, A. D'Amico, J. Richie, et al, Gene expression correlates of clinical prostate cancer behavior, Cancer Cell 1 (2) (2002) 203–209.
[107] L. Song, A. Smola, A. Gretton, J. Bedo, K. Borgwardt, Feature selection via dependence maximization, J. Mach. Learning Res. 98888 (2012) 1393–1434.
[108] Q. Song, J. Ni, G. Wang, A fast clustering-based feature subset selection algorithm for high-dimensional data, IEEE Trans. Knowl. Data Eng. 25 (1) (2013) 1–14.
[109] A. Spira, J. Beane, V. Shah, K. Steiling, G. Liu, F. Schembri, S. Gilman, Y. Dumas, P. Calner, P. Sebastiani, et al, Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer, Nat. Med. 13 (3) (2007) 361–366.
[110] A. Statnikov, C. Aliferis, I. Tsamardinos, Gems: Gene Expression Model Selector. <http://www.gems-system.org> (accessed January, 2014).
[111] J. Staunton, D. Slonim, H. Coller, P. Tamayo, M. Angelo, J. Park, U. Scherf, J. Lee, W. Reinhold, J. Weinstein, et al, Chemosensitivity prediction by transcriptional profiling, Proc. Nat. Acad. Sci. 98 (19) (2001) 10787–10792.
[112] R. Stienstra, F. Saudale, C. Duval, S. Keshtkar, J. Groener, N. van Rooijen, B. Staels, S. Kersten, M. Mnller, Kupffer cells promote hepatic steatosis via interleukin-1beta-dependent suppression of peroxisome proliferator-activated receptor alpha activity, Hepatology 51 (2) (2010) 511–522.
[113] S. Student, K. Fujarewicz, Stable feature selection and classification algorithms for multiclass microarray data, Biol. Direct 7 (1) (2012) 33.
[114] A. Su, J. Welsh, L. Sapinoso, S. Kern, P. Dimitrov, H. Lapp, P. Schultz, S. Powell, C. Moskaluk, H. Frierson Jr., et al, Molecular classification of human carcinomas by use of gene expression signatures, Cancer Res. 61 (20) (2001) 7388–7393.
[115] L. Sun, A. Hui, Q. Su, A. Vortmeyer, Y. Kotliarov, S. Pastorino, A. Passaniti, J. Menon, J. Walling, R. Bailey, et al, Neuronal and glioma-derived stem cell factor induces angiogenesis within the brain, Cancer Cell 9 (4) (2006) 287–300.
[116] Y. Sun, A.K. Wong, M.S. Kamel, Classification of imbalanced data: a review, Int. J. Pattern Recognit. Artif. Intell. 23 (04) (2009) 687–719.
[117] E. Tian, F. Zhan, R. Walker, E. Rasmussen, Y. Ma, B. Barlogie, J. Shaughnessy Jr., The role of the wnt-signaling antagonist dkk1 in the development of osteolytic lesions in multiple myeloma, New England J. Med. 349 (26) (2003) 2483–2494.
[118] L. van't Veer, H. Dai, M. Van De Vijver, Y. He, A. Hart, M. Mao, H. Peterse, K. van der Kooy, M. Marton, A. Witteveen, et al, Gene expression profiling predicts clinical outcome of breast cancer, Nature 415 (6871) (2002) 530–536.
[119] G. Wang, Q. Song, B. Xu, Y. Zhou, Selecting feature subset for high dimensional data via the propositional foil rules, Pattern Recognition 46 (1) (2013) 199–214.
[120] M. Wanderley, V. Gardeux, R. Natowicz, A. Braga, Ga-kde-bayes: an evolutionary wrapper method based on non-parametric density estimation applied to bioinformatics problems, in: 21st European Symposium on Artificial Neural Networks-ESANN, 2013, pp. 155–160.
[121] I. Wang, Y. Tetko, M. Hall, E. Frank, A. Facius, K. Mayer, H. Mewes, Gene selection from microarray data for cancer classification: a machine learning approach, Comput. Biol. Chem. 29 (1) (2005) 37–46.
[122] J. Wang, L. Wu, J. Kong, Y. Li, B. Zhang, Maximum weight and minimum redundancy: a novel framework for feature subset selection, Pattern Recognit. 46 (6) (2013) 1616–1627.

[123] M. West, C. Blanchette, H. Dressman, E. Huang, S. Ishida, R. Spang, H. Zuzan, J. Olson, J. Marks, J. Nevins, Predicting the clinical status of human breast cancer by using gene expression profiles, Proc. Nat. Acad. Sci. 98 (20) (2001) 11462–11467.
[124] F. Yang, K. Mao, Robust feature selection for microarray data based on multicriterion fusion, IEEE/ACM Trans. Comput. Biol. Bioinformatics 8 (4) (2011) 1080–1092.
[125] Y. Ye, Q. Wu, J. Zhexue Huang, M. Ng, X. Li, Stratified sampling for feature subspace selection in random forests for high dimensional data, Pattern Recognit. 46 (3) (2013) 769–787.
[126] K. Yeung, R. Bumgarner, et al, Multiclass classification of microarray data with repeated measurements: application to cancer, Genome Biol. 4 (12) (2003). 83-83.
[127] L. Yu, H. Liu, Feature selection for high-dimensional data: a fast correlation-based filter solution, in: Machine Learning-International Workshop then Conference-, vol. 20, 2003, pp. 856.
[128] Z. Zhao, H. Liu, Searching for interacting features, in: Proceedings of the 20th International Joint Conference on Artifical Intelligence, Morgan Kaufmann Publishers Inc., 2007, pp. 1156–1161.