BIOINFORMATICS REVIEW

Vol. 23 no. 19 2007, pages 2507–2517
doi:10.1093/bioinformatics/btm344

Gene expression

A review of feature selection techniques in bioinformatics

Yvan Saeys[1,*], Iñaki Inza[2] and Pedro Larrañaga[2]

[1]Department of Plant Systems Biology, VIB, B-9052 Ghent, Belgium and Bioinformatics and Evolutionary Genomics group, Department of Molecular Genetics, Ghent University, B-9052 Ghent, Belgium and [2]Department of Computer Science and Artificial Intelligence, Computer Science Faculty, University of the Basque Country, Paseo Manuel de Lardizabal 1, 20018 Donostia - San Sebastián, Spain

Received on April 17, 2007; revised on June 11, 2007; accepted on June 25, 2007

Advance Access publication August 24, 2007

Associate Editor: Jonathan Wren

ABSTRACT

Feature selection techniques have become an apparent need in many bioinformatics applications. In addition to the large pool of techniques that have already been developed in the machine learning and data mining fields, specific applications in bioinformatics have led to a wealth of newly proposed techniques.

In this article, we make the interested reader aware of the possibilities of feature selection, providing a basic taxonomy of feature selection techniques, and discussing their use, variety and potential in a number of both common as well as upcoming bioinformatics applications.

Contact: yvan.saeys@psb.ugent.be

Supplementary information: http://bioinformatics.psb.ugent.be/supplementary_data/yvsae/fsreview

1 INTRODUCTION

During the last decade, the motivation for applying feature selection (FS) techniques in bioinformatics has shifted from being an illustrative example to becoming a real prerequisite for model building. In particular, the high dimensional nature of many modelling tasks in bioinformatics, going from sequence analysis over microarray analysis to spectral analyses and literature mining has given rise to a wealth of feature selection techniques being presented in the field.

In this review, we focus on the application of feature selection techniques. In contrast to other dimensionality reduction techniques like those based on projection (e.g. principal component analysis) or compression (e.g. using information theory), feature selection techniques do not alter the original representation of the variables, but merely select a subset of them. Thus, they preserve the original semantics of the variables, hence, offering the advantage of interpretability by a domain expert.

While feature selection can be applied to both supervised and unsupervised learning, we focus here on the problem of supervised learning (classification), where the class labels are known beforehand. The interesting topic of feature selection for unsupervised learning (clustering) is a more complex issue, and research into this field is recently getting more attention in several communities (Liu and Yu, 2005; Varshavsky *et al.*, 2006).

The main aim of this review is to make practitioners aware of the benefits, and in some cases even the necessity of applying feature selection techniques. Therefore, we provide an overview of the different feature selection techniques for classification: we illustrate them by reviewing the most important application fields in the bioinformatics domain, highlighting the efforts done by the bioinformatics community in developing novel and adapted procedures. Finally, we also point the interested reader to some useful data mining and bioinformatics software packages that can be used for feature selection.

2 FEATURE SELECTION TECHNIQUES

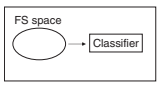
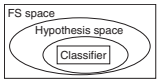
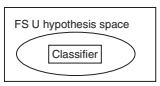
As many pattern recognition techniques were originally not designed to cope with large amounts of irrelevant features, combining them with FS techniques has become a necessity in many applications (Guyon and Elisseeff, 2003; Liu and Motoda, 1998; Liu and Yu, 2005). The objectives of feature selection are manifold, the most important ones being: (a) to avoid overfitting and improve model performance, i.e. prediction performance in the case of supervised classification and better cluster detection in the case of clustering, (b) to provide faster and more cost-effective models and (c) to gain a deeper insight into the underlying processes that generated the data. However, the advantages of feature selection techniques come at a certain price, as the search for a subset of relevant features introduces an additional layer of complexity in the modelling task. Instead of just optimizing the parameters of the model for the full feature subset, we now need to find the optimal model parameters for the optimal feature subset, as there is no guarantee that the optimal parameters for the full feature set are equally optimal for the optimal feature subset (Daelemans *et al.*, 2003). As a result, the search in the model hypothesis space is augmented by another dimension: the one of finding the optimal subset of relevant features. Feature selection techniques differ from each other in the way they

*To whom correspondence should be addressed.

© The Author 2007. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org

Downloaded from http://bioinformatics.oxfordjournals.org/ at milner library illinois state university on December 4, 2012

Table 1. A taxonomy of feature selection techniques. For each feature selection type, we highlight a set of characteristics which can guide the choice for a technique suited to the goals and resources of practitioners in the field

Model search	Advantages	Disadvantages	Examples
Filter	Univariate		
	Fast Scalable Independent of the classifier	Ignores feature dependencies Ignores interaction with the classifier	χ^2 Euclidean distance *i*-test Information gain, Gain ratio (Ben-Bassat, 1982)
	Multivariate		
	Models feature dependencies Independent of the classifier Better computational complexity than wrapper methods	Slower than univariate techniques Less scalable than univariate techniques Ignores interaction with the classifier	Correlation-based feature selection (CFS) (Hall, 1999) Markov blanket filter (MBF) (Koller and Sahami, 1996) Fast correlation-based feature selection (FCBF) (Yu and Liu, 2004)
Wrapper	Deterministic		
	Simple Interacts with the classifier Models feature dependencies Less computationally intensive than randomized methods	Risk of over fitting More prone than randomized algorithms to getting stuck in a local optimum (greedy search) Classifier dependent selection	Sequential forward selection (SFS) (Kittler, 1978) Sequential backward elimination (SBE) (Kittler, 1978) Plus *q* take-away *r* (Ferri *et al*., 1994) Beam search (Siedelecky and Sklansky, 1988)
	Randomized		
	Less prone to local optima Interacts with the classifier Models feature dependencies	Computationally intensive Classifier dependent selection Higher risk of overfitting than deterministic algorithms	Simulated annealing Randomized hill climbing (Skalak, 1994) Genetic algorithms (Holland, 1975) Estimation of distribution algorithms (Inza *et al*., 2000)
Embedded	Interacts with the classifier Better computational complexity than wrapper methods Models feature dependencies	Classifier dependent selection	Decision trees Weighted naive Bayes (Duda *et al*., 2001) Feature selection using the weight vector of SVM (Guyon *et al*., 2002; Weston *et al*., 2003)
			

incorporate this search in the added space of feature subsets in the model selection.

In the context of classification, feature selection techniques can be organized into three categories, depending on how they combine the feature selection search with the construction of the classification model: filter methods, wrapper methods and embedded methods. Table 1 provides a common taxonomy of feature selection methods, showing for each technique the most prominent advantages and disadvantages, as well as some examples of the most influential techniques.

Filter techniques assess the relevance of features by looking only at the intrinsic properties of the data. In most cases a feature relevance score is calculated, and low-scoring features are removed. Afterwards, this subset of features is presented as input to the classification algorithm. Advantages of filter techniques are that they easily scale to very high-dimensional datasets, they are computationally simple and fast, and they are independent of the classification algorithm. As a result, feature selection needs to be performed only once, and then different classifiers can be evaluated.

Downloaded from http://bioinformatics.oxfordjournals.org/ at milner library illinois state university on December 4, 2012

A common disadvantage of filter methods is that they ignore the interaction with the classifier (the search in the feature subset space is separated from the search in the hypothesis space), and that most proposed techniques are univariate. This means that each feature is considered separately, thereby ignoring feature dependencies, which may lead to worse classification performance when compared to other types of feature selection techniques. In order to overcome the problem of ignoring feature dependencies, a number of multivariate filter techniques were introduced, aiming at the incorporation of feature dependencies to some degree.

Whereas filter techniques treat the problem of finding a good feature subset independently of the model selection step, *wrapper methods* embed the model hypothesis search within the feature subset search. In this setup, a search procedure in the space of possible feature subsets is defined, and various subsets of features are generated and evaluated. The evaluation of a specific subset of features is obtained by training and testing a specific classification model, rendering this approach tailored to a specific classification algorithm. To search the space of all feature subsets, a search algorithm is then 'wrapped' around the classification model. However, as the space of feature subsets grows exponentially with the number of features, heuristic search methods are used to guide the search for an optimal subset. These search methods can be divided in two classes: deterministic and randomized search algorithms. Advantages of wrapper approaches include the interaction between feature subset search and model selection, and the ability to take into account feature dependencies. A common drawback of these techniques is that they have a higher risk of overfitting than filter techniques and are very computationally intensive, especially if building the classifier has a high computational cost.

In a third class of feature selection techniques, termed *embedded techniques*, the search for an optimal subset of features is built into the classifier construction, and can be seen as a search in the combined space of feature subsets and hypotheses. Just like wrapper approaches, embedded approaches are thus specific to a given learning algorithm. Embedded methods have the advantage that they include the interaction with the classification model, while at the same time being far less computationally intensive than wrapper methods.

3 APPLICATIONS IN BIOINFORMATICS

3.1 Feature selection for sequence analysis

Sequence analysis has a long-standing tradition in bioinformatics. In the context of feature selection, two types of problems can be distinguished: content and signal analysis. Content analysis focuses on the broad characteristics of a sequence, such as tendency to code for proteins or fulfillment of a certain biological function. Signal analysis on the other hand focuses on the identification of important motifs in the sequence, such as gene structural elements or regulatory elements.

Apart from the basic features that just represent the nucleotide or amino acid at each position in a sequence, many other features, such as higher order combinations of these building blocks (e.g. k-mer patterns) can be derived, their number growing exponentially with the pattern length k. As many of them will be irrelevant or redundant, feature selection techniques are then applied to focus on the subset of relevant variables.

3.1.1 Content analysis The prediction of subsequences that code for proteins (coding potential prediction) has been a focus of interest since the early days of bioinformatics. Because many features can be extracted from a sequence, and most dependencies occur between adjacent positions, many variations of Markov models were developed. To deal with the high amount of possible features, and the often limited amount of samples, (Salzberg *et al.*, 1998) introduced the interpolated Markov model (IMM), which used interpolation between different orders of the Markov model to deal with small sample sizes, and a filter method (χ^2) to select only relevant features. In further work, (Delcher *et al.*, 1999) extended the IMM framework to also deal with non-adjacent feature dependencies, resulting in the interpolated context model (ICM), which crosses a Bayesian decision tree with a filter method (χ^2) to assess feature relevance. Recently, the avenue of FS techniques for coding potential prediction was further pursued by (Saeys *et al.*, 2007), who combined different measures of coding potential prediction, and then used the Markov blanket multivariate filter approach (MBF) to retain only the relevant ones.

A second class of techniques focuses on the prediction of protein function from sequence. The early work of Chuzhanova *et al.* (1998), who combined a genetic algorithm in combination with the Gamma test to score feature subsets for classification of large subunits of rRNA, inspired researchers to use FS techniques to focus on important subsets of amino acids that relate to the protein's functional class (Al-Shahib *et al.*, 2005). An interesting technique is described in Zavaljevsky *et al.* (2002), using selective kernel scaling for support vector machines (SVM) as a way to asses feature weights, and subsequently remove features with low weights.

The use of FS techniques in the domain of sequence analysis is also emerging in a number of more recent applications, such as the recognition of promoter regions (Conilione and Wang, 2005), and the prediction of microRNA targets (Kim *et al.*, 2006).

3.1.2 Signal analysis Many sequence analysis methodologies involve the recognition of short, more or less conserved signals in the sequence, representing mainly binding sites for various proteins or protein complexes. A common approach to find regulatory motifs, is to relate motifs to gene expression levels using a regression approach. Feature selection can then be used to search for the motifs that maximize the fit to the regression model (Keles *et al.*, 2002; Tadesse *et al.*, 2004). In Sinha (2003), a classification approach is chosen to find discriminative motifs. The method is inspired by Ben-Dor *et al.* (2000) who use the threshold number of misclassification (TNoM, see further in the section on microarray analysis) to score genes for relevance to tissue classification. From the TNoM score, a P-value is calculated that represents the significance of each motif. Motifs are then sorted according to their P-value.

Downloaded from http://bioinformatics.oxfordjournals.org/ at milner library illinois state university on December 4, 2012

Table 2. Key references for each type of feature selection technique in the microarray domain

Filter methods			Wrapper methods	Embedded methods
Univariate		Multivariate		
Parametric	Model-free			
t-test (Jafari and Azuaje, 2006)	Wilcoxon rank sum (Thomas *et al*., 2001)	Bivariate (Bø and Jonassen, 2002)	Sequential search (Inza *et al*., 2004; Xiong *et al*., 2001)	Random forest (Díaz-Uriarte and Alvarez de Andrés, 2006; Jiang *et al*., 2004)
ANOVA (Jafari and Azuaje, 2006)	BSS/WSS (Dudoit *et al*., 2002)	CFS (Wang *et al*., 2005; Yeoh et al., 2002)	Genetic algorithms (Jirapech-Umpai and Aitken, 2005; Li *et al*., 2001; Ooi and Tan, 2003)	Weight vector of SVM (Guyon *et al*., 2002)
Bayesian (Baldi and Long, 2001; Fox and Dimmic, 2006)	Rank products (Breitling *et al*., 2004)	MRMR (Ding and Peng, 2003)	Estimation of distribution algorithms (Blanco *et al*., 2004)	Weights of logistic regression (Ma and Huang, 2005)
Regression (Thomas *et al*., 2001)	Random permutations (Efron *et al*., 2001; Pan, 2003; Park *et al*., 2001; Tusher *et al*., 2001)	USC (Yeung and Bumgarner, 2003)		
		Markov blanket (Gevaert *et al*., 2006; Mamitsuka, 2006; Xing *et al*., 2001)		
Gamma (Newton *et al*., 2001)	TNoM (Ben-Dor *et al*., 2000)			

Another line of research is performed in the context of the gene prediction setting, where structural elements such as the translation initiation site (TIS) and splice sites are modelled as specific classification problems. The problem of feature selection for structural element recognition was pioneered in Degroeve *et al*. (2002) for the problem of splice site prediction, combining a sequential backward method together with an embedded SVM evaluation criterion to assess feature relevance. In Saeys *et al*. (2004), an estimation of distribution algorithm (EDA, a generalization of genetic algorithms) was used to gain more insight in the relevant features for splice site prediction. Similarly, the prediction of TIS is a suitable problem to apply feature selection techniques. In Liu *et al*. (2004), the authors demonstrate the advantages of using feature selection for this problem, using the feature-class entropy as a filter measure to remove irrelevant features.

In future research, FS techniques can be expected to be useful for a number of challenging prediction tasks, such as identifying relevant features related to alternative splice sites and alternative TIS.

3.2 Feature selection for microarray analysis

During the last decade, the advent of microarray datasets stimulated a new line of research in bioinformatics. Microarray data pose a great challenge for computational techniques, because of their large dimensionality (up to several tens of thousands of genes) and their small sample sizes (Somorjai *et al*., 2003). Furthermore, additional experimental complications like noise and variability render the analysis of microarray data an exciting domain.

In order to deal with these particular characteristics of microarray data, the obvious need for dimension reduction techniques was realized (Alon *et al*., 1999; Ben-Dor *et al*., 2000; Golub *et al*., 1999; Ross *et al*., 2000), and soon their application became a *de facto* standard in the field. Whereas in 2001, the field of microarray analysis was still claimed to be in its infancy (Efron *et al*., 2001), a considerable and valuable effort has since been done to contribute new and adapt known FS methodologies (Jafari and Azuaje, 2006). A general overview of the most influential techniques, organized according to the general FS taxonomy of Section 2, is shown in Table 2.

3.2.1 The univariate filter paradigm: simple yet efficient

Because of the high dimensionality of most microarray analyses, fast and efficient FS techniques such as univariate filter methods have attracted most attention. The prevalence of these univariate techniques has dominated the field, and up to now comparative evaluations of different classification and FS techniques over DNA microarray datasets only focused on the univariate case (Dudoit *et al*., 2002; Lee *et al*., 2005; Li *et al*., 2004; Statnikov *et al*., 2005). This domination of the univariate approach can be explained by a number of reasons:

- the output provided by univariate feature rankings is intuitive and easy to understand;

Downloaded from http://bioinformatics.oxfordjournals.org/ at milner library illinois state university on December 4, 2012

- the gene ranking output could fulfill the objectives and expectations that bio-domain experts have when wanting to subsequently validate the result by laboratory techniques or in order to explore literature searches. The experts could not feel the need for selection techniques that take into account gene interactions;
- the possible unawareness of subgroups of gene expression domain experts about the existence of data analysis techniques to select genes in a multivariate way;
- the extra computation time needed by multivariate gene selection techniques.

Some of the simplest heuristics for the identification of differentially expressed genes include setting a threshold on the observed fold-change differences in gene expression between the states under study, and the detection of the threshold point in each gene that minimizes the number of training sample misclassification (threshold number of misclassification, TNoM (Ben-Dor *et al*., 2000)). However, a wide range of new or adapted univariate feature ranking techniques has since then been developed. These techniques can be divided into two classes: parametric and model-free methods (see Table 2).

Parametric methods assume a given distribution from which the samples (observations) have been generated. The two sample *t*-test and ANOVA are among the most widely used techniques in microarray studies, although the usage of their basic form, possibly without justification of their main assumptions, is not advisable (Jafari and Azuaje, 2006). Modifications of the standard *t*-test to better deal with the small sample size and inherent noise of gene expression datasets include a number of *t*- or *t*-test like statistics (differing primarily in the way the variance is estimated) and a number of Bayesian frameworks (Baldi and Long, 2001; Fox and Dimmic, 2006). Although Gaussian assumptions have dominated the field, other types of parametrical approaches can also be found in the literature, such as regression modelling approaches (Thomas *et al*., 2001) and Gamma distribution models (Newton *et al*., 2001).

Due to the uncertainty about the true underlying distribution of many gene expression scenarios, and the difficulties to validate distributional assumptions because of small sample sizes, non-parametric or model-free methods have been widely proposed as an attractive alternative to make less stringent distributional assumptions (Troyanskaya *et al*., 2002). Many model-free metrics, frequently borrowed from the statistics field, have demonstrated their usefulness in many gene expression studies, including the Wilcoxon rank-sum test (Thomas *et al*., 2001), the between-within classes sum of squares (BSS/WSS) (Dudoit *et al*., 2002) and the rank products method (Breitling *et al*., 2004).

A specific class of model-free methods estimates the reference distribution of the statistic using random permutations of the data, allowing the computation of a model-free version of the associated parametric tests. These techniques have emerged as a solid alternative to deal with the specificities of DNA microarray data, and do not depend on strong parametric assumptions (Efron *et al*., 2001; Pan, 2003; Park *et al*., 2001; Tusher *et al*., 2001). Their permutation principle partly alleviates the problem of small sample sizes in microarray studies, enhancing the robustness against outliers.

We also mention promising types of non-parametric metrics which, instead of trying to identify differentially expressed genes at the whole population level (e.g. comparison of sample means), are able to capture genes which are significantly disregulated in only a subset of samples (Lyons-Weiler *et al*., 2004; Pavlidis and Poirazi, 2006). These types of methods offer a more patient specific approach for the identification of markers, and can select genes exhibiting complex patterns that are missed by metrics that work under the classical comparison of two prelabelled phenotypic groups. In addition, we also point out the importance of procedures for controlling the different types of errors that arise in this complex multiple testing scenario of thousands of genes (Dudoit *et al*., 2003; Ploner *et al*., 2006; Pounds and Cheng, 2004; Storey, 2002), with a special focus on contributions for controlling the false discovery rate (FDR).

3.2.2 Towards more advanced models: the multivariate paradigm for filter, wrapper and embedded techniques
Univariate selection methods have certain restrictions and may lead to less accurate classifiers by, e.g. not taking into account gene–gene interactions. Thus, researchers have proposed techniques that try to capture these correlations between genes.

The application of multivariate filter methods ranges from simple bivariate interactions (Bø and Jonassen, 2002) towards more advanced solutions exploring higher order interactions, such as correlation-based feature selection (CFS) (Wang *et al*., 2005; Yeoh *et al*., 2002) and several variants of the Markov blanket filter method (Gevaert *et al*., 2006; Mamitsuka, 2006; Xing *et al*., 2001). The Minimum Redundancy-Maximum Relevance (MRMR) (Ding and Peng, 2003) and Uncorrelated Shrunken Centroid (USC) (Yeung and Bumgarner, 2003) algorithms are two other solid multivariate filter procedures, highlighting the advantage of using multivariate methods over univariate procedures in the gene expression domain.

Feature selection using wrapper or embedded methods offers an alternative way to perform a multivariate gene subset selection, incorporating the classifier's bias into the search and thus offering an opportunity to construct more accurate classifiers. In the context of microarray analysis, most wrapper methods use population-based, randomized search heuristics (Blanco *et al*., 2004; Jirapech-Umpai and Aitken, 2005; Li *et al*., 2001; Ooi and Tan, 2003), although also a few examples use sequential search techniques (Inza *et al*., 2004; Xiong *et al*., 2001). An interesting hybrid filter-wrapper approach is introduced in (Ruiz *et al*., 2006), crossing a univariately pre-ordered gene ranking with an incrementally augmenting wrapper method.

Another characteristic of any wrapper procedure concerns the scoring function used to evaluate each gene subset found. As the 0–1 accuracy measure allows for comparison with previous works, the vast majority of papers uses this measure. However, recent proposals advocate the use of methods for the approximation of the area under the ROC curve (Ma and Huang, 2005), or the optimization of the LASSO (Least Absolute Shrinkage and Selection Operator) model (Ghosh and Chinnaiyan, 2005). ROC curves certainly provide an interesting

Downloaded from http://bioinformatics.oxfordjournals.org/ at milner library illinois state university on December 4, 2012

Table 3. . Key references for each type of feature selection technique in the domain of mass spectrometry

Filter	Univariate		Multivariate
	Parametric	Model-free	
	t-test (Liu *et al*., 2002; Wu *et al*., 2003) *F*-test (Bhanot *et al*., 2006)	Peak Probability Contrast (Tibshirani *et al*., 2004) Kolmogorov-Smirnov test (Yu *et al*., 2005)	CFS (Liu *et al*., 2002) Relief-F (Prados *et al*., 2004)
Wrapper	Genetic algorithms (Li *et al*., 2004; Petricoin *et al*., 2002) Nature inspired (Ressom *et al*., 2005, 2007)		
Embedded	Random forest/decision tree (Geurts *et al*., 2005; Wu *et al*., 2003) Weight vector of SVM (Jong *et al*., 2004; Prados *et al*., 2004; Zhang *et al*., 2006) Neural network (Ball *et al*., 2002)		

evaluation measure, especially suited to the demand for screening different types of errors in many biomedical scenarios.

The embedded capacity of several classifiers to discard input features and thus propose a subset of discriminative genes, has been exploited by several authors. Examples include the use of random forests (a classifier that combines many single decision trees) in an embedded way to calculate the importance of each gene (Díaz-Uriarte and Alvarez de Andrés, 2006; Jiang *et al*., 2004). Another line of embedded FS techniques uses the weights of each feature in linear classifiers, such as SVMs (Guyon *et al*., 2002) and logistic regression (Ma and Huang, 2005). These weights are used to reflect the relevance of each gene in a multivariate way, and thus allow for the removal of genes with very small weights.

Partially due to the higher computational complexity of wrapper and to a lesser degree embedded approaches, these techniques have not received as much interest as filter proposals. However, an advisable practice is to pre-reduce the search space using a univariate filter method, and only then apply wrapper or embedded methods, hence fitting the computation time to the available resources.

3.3 Mass spectra analysis

Mass spectrometry technology (MS) is emerging as a new and attractive framework for disease diagnosis and protein-based biomarker profiling (Petricoin and Liotta, 2003). A mass spectrum sample is characterized by thousands of different mass/charge (m/z) ratios on the x-axis, each with their corresponding signal intensity value on the y-axis. A typical MALDI-TOF low-resolution proteomic profile can contain up to 15 500 data points in the spectrum between 500 and 20 000 m/z, and the number of points even grows using higher resolution instruments.

For data mining and bioinformatics purposes, it can initially be assumed that each m/z ratio represents a distinct variable whose value is the intensity. As Somorjai *et al*. (2003) explain, the data analysis step is severely constrained by both high-dimensional input spaces and their inherent sparseness, just as it is the case with gene expression datasets. Although the amount of publications on mass spectrometry based data mining is not comparable to the level of maturity reached in the microarray analysis domain, an interesting collection of methods has been presented in the last 4–5 years (see Hilario *et al*., 2006; Shin and Markey, 2006 for recent reviews) since the pioneering work of Petricoin *et al*. (2002).

Starting from the raw data, and after an initial step to reduce noise and normalize the spectra from different samples (Coombes *et al*., 2007), the following crucial step is to extract the variables that will constitute the initial pool of candidate discriminative features. Some studies employ the simplest approach of considering every measured value as a predictive feature, thus applying FS techniques over initial huge pools of about 15 000 variables (Li *et al*., 2004; Petricoin *et al*., 2002), up to around 100 000 variables (Ball *et al*., 2002). On the other hand, a great deal of the current studies performs aggressive feature extraction procedures using elaborated peak detection and alignment techniques (see Coombes *et al*., 2007; Hilario *et al*., 2006; Shin and Markey, 2006 for a detailed description of these techniques). These procedures tend to seed the dimensionality from which supervised FS techniques will start their work in less than 500 variables (Bhanot *et al*., 2006; Ressom *et al*., 2007; Tibshirani *et al*., 2004). A feature extraction step is thus advisable to set the computational costs of many FS techniques to a feasible size in these MS scenarios. Table 3 presents an overview of FS techniques used in the domain of mass spectrometry. Similar to the domain of microarray analysis, univariate filter techniques seem to be the most common techniques used, although the use of embedded techniques is certainly emerging as an alternative. Although the *t*-test maintains a high level of popularity (Liu *et al*., 2002; Wu *et al*., 2003), other parametric measures such as *F*-test (Bhanot *et al*., 2006), and a notable variety of non-parametric scores (Tibshirani *et al*., 2004; Yu *et al*., 2005) have also been used in several MS studies. Multivariate filter techniques on the other hand, are still somewhat underrepresented (Liu *et al*., 2002; Prados *et al*., 2004).

Wrapper approaches have demonstrated their usefulness in MS studies by a group of influential works. Different types of population-based randomized heuristics are used as search engines in the major part of these papers: genetic algorithms (Li *et al*., 2004; Petricoin *et al*., 2002), particle swarm

Downloaded from http://bioinformatics.oxfordjournals.org/ at milner library illinois state university on December 4, 2012

optimization (Ressom *et al*., 2005) and ant colony procedures (Ressom *et al*., 2007). It is worth noting that while the first two references start the search procedure in ~15 000 dimensions by considering each m/z ratio as an initial predictive feature, aggressive peak detection and alignment processes reduce the initial dimension to about 300 variables in the last two references (Ressom *et al*., 2005; Ressom *et al*., 2007).

An increasing number of papers uses the embedded capacity of several classifiers to discard input features. Variations of the popular method originally proposed for gene expression domains by Guyon *et al*. (2002), using the weights of the variables in the SVM-formulation to discard features with small weights, have been broadly and successfully applied in the MS domain (Jong *et al*., 2004; Prados *et al*., 2004; Zhang *et al*., 2006). Based on a similar framework, the weights of the input masses in a neural network classifier have been used to rank the features' importance in Ball *et al*. (2002). The embedded capacity of random forests (Wu *et al*., 2003) and other types of decision tree-based algorithms (Geurts *et al*., 2005) constitutes an alternative embedded FS strategy.

4 DEALING WITH SMALL SAMPLE DOMAINS

Small sample sizes, and their inherent risk of imprecision and overfitting, pose a great challenge for many modelling problems in bioinformatics (Braga-Neto and Dougherty, 2004; Molinaro *et al*., 2005; Sima and Dougherty, 2006). In the context of feature selection, two initiatives have emerged in response to this novel experimental situation: the use of adequate evaluation criteria, and the use of stable and robust feature selection models.

4.1 Adequate evaluation criteria

Several papers have warned about the substantial number of applications not performing an independent and honest validation of the reported accuracy percentages (Ambroise and McLachlan, 2002; Statnikov *et al*., 2005; Somorjai *et al*., 2003). In such cases, authors often select a discriminative subset of features using the whole dataset. The accuracy of the final classification model is estimated using this subset, thus testing the discrimination rule on samples that were already used to propose the final subset of features. We feel that the need for an external feature selection process in training the classification rule at each stage of the accuracy estimation procedure is gaining space in the bioinformatics community practices.

Furthermore, novel predictive accuracy estimation methods with promising characteristics, such as bolstered error estimation (Sima *et al*., 2005), have emerged to deal with the specificities of small sample domains.

4.2 Ensemble feature selection approaches

Instead of choosing one particular FS method, and accepting its outcome as the final subset, different FS methods can be combined using *ensemble FS approaches*. Based on the evidence that there is often not a single universally optimal feature selection technique (Yang *et al*., 2005), and due to the possible existence of more than one subset of features that discriminates the data equally well (Yeung *et al*., 2005), model combination approaches such as boosting have been adapted to improve the robustness and stability of final, discriminative methods (Ben-Dor *et al*., 2000; Dudoit *et al*., 2002).

Novel ensemble techniques in the microarray and mass spectrometry domains include averaging over multiple single feature subsets (Levner, 2005; Li and Yang, 2002), integrating a collection of univariate differential gene expression purpose statistics via a distance synthesis scheme (Yang *et al*., 2005), using different runs of a genetic algorithm to asses relative importancies of each feature (Li *et al*., 2001, 2004), computing the Kolmogorov–Smirnov test in different bootstrap samples to assign a probability of being selected to each peak (Yu and Chen, 2005), and a number of Bayesian averaging approaches (Lee *et al*., 2003; Yeung *et al*., 2005). Furthermore, methods based on a collection of decision trees (e.g. random forests) can be used in an ensemble FS way to assess the relevance of each feature (Díaz-Uriarte and Alvarez de Andrés, 2006; Geurts *et al*., 2005; Jiang *et al*., 2004; Wu *et al*., 2003).

Although the use of ensemble approaches requires additional computational resources, we would like to point out that they offer an advisable framework to deal with small sample domains, provided the extra computational resources are affordable.

5 FEATURE SELECTION IN UPCOMING DOMAINS

5.1 Single nucleotide polymorphism analysis

Single nucleotide polymorphisms (SNPs) are mutations at a single nucleotide position that occurred during evolution and were passed on through heredity, accounting for most of the genetic variation among different individuals. SNPs are at the forefront of many disease-gene association studies, their number being estimated at about 7 million in the human genome (Kruglyak and Nickerson, 2001). Thus, selecting a subset of SNPs that is sufficiently informative but still small enough to reduce the genotyping overhead is an important step towards disease-gene association. Typically, the number of SNPs considered is not higher than tens of thousands with sample sizes of about 100.

Several computational methods for htSNP selection (haplotype SNPs; a set of SNPs located on one chromosome) have been proposed in the past few years. One approach is based on the hypothesis that the human genome can be viewed as a set of discrete blocks that only share a very small set of common haplotypes (Daly *et al*., 2001). This approach aims to identify a subset of SNPs that can either distinguish all the common haplotypes (Gabriel *et al*., 2002), or at least explain a certain percentage of them. Another common htSNP selection approach is based on pairwise associations of SNPs, and tries to select a set of htSNPs such that each of the SNPs on a haplotype is highly associated with one of the htSNPs (Carlson *et al*., 2004). A third approach considers htSNPs as a subset of all SNPs, from which the remaining SNPs can be reconstructed (Halperin *et al*., 2005; Lee and Shatkay, 2006; Lin and Altman, 2004). The idea is to select htSNPs based on how well they predict the remaining set of the unselected SNPs.

Downloaded from http://bioinformatics.oxfordjournals.org/ at milner library illinois state university on December 4, 2012

Table 4. Software for feature selection

		General purpose FS software	
WEKA	Java	Witten and Frank (2005)	http://www.cs.waikato.ac.nz/ml/weka
Fast Correlation Based Filter	Java	Yu and Liu (2004)	http://www.public.asu.edu/~huanliu/FCBF/FCBFsoftware.html
Feature Selection Book	Ansi C	Liu and Motoda (1998)	http://www.public.asu.edu/~huanliu/Fsbook
MLC++	C++	Kohavi *et al.* (1996)	http://www.sgi.com/tech/mlc
Spider	Matlab	–	http://www.kyb.tuebingen.mpg.de/bs/people/spider
SVM and Kernel Methods Matlab Toolbox	Matlab	Canu *et al.* (2003)	http://asi.insa-rouen.fr/~arakotom/toolbox/index
		Microarray analysis FS software	
SAM	R, Excel	Tusher *et al.* (2001)	http://www-stat.stanford.edu/~tibs/SAM/
GALGO	R	Trevino and Falciani (2006)	http://www.bip.bham.ac.uk/bioinf/galgo.html
PCP	C, C++	Buturovic (2005)	http://pcp.sourceforge.net
GA-KNN	C	Li *et al.* (2001)	http://dir.niehs.nih.gov/microarray/datamining/
Rankgene	C	Su *et al.* (2003)	http://genomics10.bu.edu/yangsu/rankgene/
EDGE	R	Leek *et al.* (2006)	http://www.biostat.washington.edu/software/jstorey/edge/
GEPAS-Prophet	Perl, C	Medina *et al.* (2007)	http://prophet.bioinfo.cipf.es/
DEDS (Bioconductor)	R	Yang *et al.* (2005)	http://www.bioconductor.org/
RankProd (Bioconductor)	R	Breitling *et al.* (2004)	http://www.bioconductor.org/
Limma (Bioconductor)	R	Smyth (2004)	http://www.bioconductor.org/
Multtest (Bioconductor)	R	Dudoit *et al.* (2003)	http://www.bioconductor.org/
Nudge (Bioconductor)	R	Dean and Raftery (2005)	http://www.bioconductor.org/
Qvalue (Bioconductor)	R	Storey (2002)	http://www.bioconductor.org/
twilight (Bioconductor)	R	Scheid and Spang (2005)	http://www.bioconductor.org/
ComparativeMarkerSelection (GenePattern)	JAVA, R	Gould *et al.* (2006)	http://www.broad.mit.edu/genepattern
		Mass spectra analysis FS software	
GA-KNN	C	Li *et al.* (2004)	http://dir.niehs.nih.gov/microarray/datamining/
R-SVM	R, C, C++	Zhang *et al.* (2006)	http://www.hsph.harvard.edu/bioinfocore/RSVMhome/R-SVM.html
		SNP analysis FS software	
CHOISS	C++, Perl	Lee and Kang (2004)	http://biochem.kaist.ac.kr/choiss.htm
MLR-tagging	C	He and Zelikovsky (2006)	http://alla.cs.gsu.ed/~software/tagging/tagging.html
WCLUSTAG	JAVA	Sham *et al.* (2007)	http://bioinfo.hku.hk/wclustag

When the haplotype structure in the target region is unknown, a widely used approach is to choose markers at regular intervals (Lee and Kang, 2004), given either the number of SNPs to choose or the desired interval. In (Li *et al.*, 2005) an ensemble approach is successfully applied to the identification of relevant SNPs for alcoholism, while Gong *et al.* (2005) propose a robust feature selection technique based on a hybrid between a genetic algorithm and an SVM. The Relief-F feature selection algorithm, in conjunction with three classification algorithms (k-NN, SVM and naive Bayes) has been proposed in Wang *et al.*, (2006). Genetic algorithms have been applied to the search of the best subset of SNPs, evaluating them with a multivariate filter (CFS), and also in a wrapper manner, with a decision tree as supervised classification paradigm (Shah and Kusiak, 2004). The multiple linear regression SNP prediction algorithm (He and Zelikovsky, 2006) predicts a complete genotype based on the values of its informative SNPs (selected with a stepwise tag selection algorithm), their positions among all SNPS, and a sample of complete genotypes. In Sham *et al.* (2007) the tag SNP selection method allows to specify variable tagging thresholds, based on correlations, for different SNPs.

5.2 Text and literature mining

Text and literature mining is emerging as a promising area for data mining in biology (Cohen and Hersch, 2005; Jensen *et al.* 2006). One important representation of text and documents is the so-called bag-of-words (BOW) representation, where each word in the text represents one variable, and its value consists of the frequency of the specific word in the text. It goes without saying that such a representation of the text may lead to very high dimensional datasets, pointing out the need for feature selection techniques.

Although the application of feature selection techniques is common in the field of text classification (see e.g. Forman, 2003 for a review), the application in the biomedical domain is still in its infancy. Some examples of FS techniques in the biomedical domain include the work of Dobrokhotov *et al.* (2003), who use the Kullback–Leibler divergence as a univariate filter method to find discriminating words in a medical annotation task, the work of Eom and Zhang (2000) who use symmetrical uncertainty (an entropy-based filter method) for identifying relevant features for protein interaction discovery, and the work of Han *et al.* (2006), which

Downloaded from http://bioinformatics.oxfordjournals.org/ at milner library illinois state university on December 4, 2012

discusses the use of feature selection for a document classification task.

It can be expected that, for tasks such as biomedical document clustering and classification, the large number of feature selection techniques that were already developed in the text mining community will be of practical use for researchers in biomedical literature mining (Cohen and Hersch, 2005).

6 FS SOFTWARE PACKAGES

In order to provide the interested reader with some pointers to existing software packages, Table 4 shows an overview of existing software implementing a variety of feature selection methods. All software packages mentioned are free for academic use, and the software is organized into four sections: general purpose FS techniques, techniques tailored to the domain of microarray analysis, techniques specific to the domain of mass spectra analysis and techniques to handle SNP selection. For each software package, the main reference, implementation language and website is shown.

In addition to these publicly available packages, we also provide a companion website as Supplementary Matrial of this work (see the Abstract section for the location). On this website, the publications are indexed according to the FS technique used, a number of keywords accompanying each reference to understand its FS methodological contributions.

7 CONCLUSIONS AND FUTURE PERSPECTIVES

In this article, we reviewed the main contributions of feature selection research in a set of well-known bioinformatics applications. Two main issues emerge as common problems in the bioinformatics domain: the large input dimensionality, and the small sample sizes. To deal with these problems, a wealth of FS techniques has been designed by researchers in bioinformatics, machine learning and data mining.

A large and fruitful effort has been performed during the last years in the adaptation and proposal of univariate filter FS techniques. In general, we observe that many researchers in the field still think that filter FS approaches are only restricted to univariate approaches. The proposal of multivariate selection algorithms can be considered as one of the most promising future lines of work for the bioinformatics community.

A second line of future research is the development of especially fitted ensemble FS approaches to enhance the robustness of the finally selected feature subsets. We feel that, in order to alleviate the actual small sample sizes of the majority of bioinformatics applications, the further development of such techniques, combined with appropriate evaluation criteria, constitutes an interesting direction for future FS research.

Other interesting opportunities for future FS research will be the extension towards upcoming bioinformatics domains, such as SNPs, text and literature mining, and the combination of heterogeneous data sources. While in these domains, the FS component is not yet as central as, e.g. in gene expression or MS areas, we believe that its application will become essential in dealing with the high-dimensional character of these applications.

To conclude, we would like to note that, in order to maintain an appropriate size of the article, we had to limit the number of referenced studies. We therefore apologize to the authors of papers that were not cited in this work.

ACKNOWLEDGEMENTS

We would like to thank the anonymous reviewers for their constructive comments, which significantly improved the quality of this review. This work was supported by BOF grant 01P10306 from Ghent University to Y.S., and the SAIOTEK and ETORTEK programs of the Basque Government and project TIN2005-03824 of the Spanish Ministry of Education and Science to I.I. and P.L.

Conflict of Interest: none declared.

REFERENCES

Alon,U. *et al.* (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Nat. Acad. Sci. USA*, **96**, 6745–6750.

Al-Shahib,A., *et al.* (2005) Feature selection and the class imbalance problem in predicting protein function from sequence. *Appl. Bioinformatics*, **4**, 195–203.

Ambroise,C. and McLachlan,G. (2002) Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc. Nat. Acad. Sci. USA*, **99**, 6562–6566.

Baldi,P. and Long,A. (2001) A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics*, **17**, 509–516.

Ball,G., *et al.* (2002) An integrated approach utilizing artificial neural networks and SELDI mass spectrometry for the classification of human tumours and rapid identification of potential biomarkers. *Bioinformatics*, **18**, 395–404.

Ben-Bassat,M. (1982) Pattern recognition and reduction of dimensionality. In Krishnaiah,P. and Kanal,L., (eds.) *Handbook of Statistics II*, Vol. 1. North-Holland, Amsterdam. pp. 773–791.

Ben-Dor,A., *et al.* (2000) Tissue classification with gene expression profiles. *J. Comput. Biol.*, **7**, 559–584.

Bhanot,G., *et al.* (2006) A robust meta classification strategy for cancer detection from MS data. *Proteomics*, **6**, 592–604.

Blanco,R., *et al.* (2004) Gene selection for cancer classification using wrapper approaches. *Int. J. Pattern Recognit. Artif. Intell.*, **18**, 1373–1390.

Bø,T. and Jonassen,I. (2002) New feature subset selection procedures for classification of expression profiles. *Genome Biol.*, **3**, research0017.1–research0017.11.

Braga-Neto,U. and Dougherty,E. (2004) Is cross-validation valid for small-sample microarray classification? *Bioinformatics*, **20**, 374–380.

Breitling,R., *et al.* (2004) Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Lett.*, **573**, 83–92.

Buturovic,L. (2005) PCP: a program for supervised classification of gene expression profiles. *Bioinformatics*, **22**, 245–247.

Canu,S., *et al.* (2003) SVM and Kernel Methods Matlab Toolbox. In *Perception Systèmes et Information*. INSA de Rouen, Rouen, France.

Carlson,C., *et al.* (2004) Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am. J. Hum. Genet.*, **74**, 106–120.

Chuzhanova,N., *et al.* (1998) Feature selection for genetic sequence classification. *Bioinformatics*, **14**, 139–143.

Cohen,A. and Hersch,W. (2005) A survey of current work in biomedical text mining. *Brief. Bioinformatics*, **6**, 57–71.

Conilione,P. and Wang,D. (2005) A comparative study on feature selection for E.coli promoter recognition. *Int. J. Inf. Technol.*, **11**, 54–66.

Coombes,K., *et al.* (2007) Pre-processing mass spectrometry data. In Dubitzky,M., *et al.* (eds.), *Fundamentals of Data Mining in Genomics and Proteomics*. Kluwer, Boston, pp. 79–99.

Daelemans,W., *et al.* (2003) Combined optimization of feature selection and algorithm parameter interaction in machine learning of language.

Downloaded from http://bioinformatics.oxfordjournals.org/ at milner library illinois state university on December 4, 2012

In *Proceedings of the 14th European Conference on Machine Learning (ECML-2003)*, pp. 84–95.

Daly,M., *et al*. (2001) High-resolution haplotype structure in the human genome. *Nat. Genet.*, **29**, 229–232.

Dean,N. and Raftery,A. (2005) Normal uniform mixture differential gene expression detection in cDNA microarrays. *BMC Bioinformatics*, **6**, 173.

Degroeve,S., *et al*. (2002) Feature subset selection for splice site prediction. *Bioinformatics*, **18** (Suppl. 2), 75–83.

Delcher,A., *et al*. (1999) Improved microbial gene identification with GLIMMER. *Nucleic Acids Res.*, **27**, 4636–4641.

Díaz-Uriarte,R. and Alvarez de Andrés,S. (2006) Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, **7**, 3.

Ding,C. and Peng,H. (2003) Minimum redundancy feature selection from microarray gene expression data. In *Proceedings of the IEEE Conference on Computational Systems Bioinformatics*, pp. 523–528.

Dobrokhotov,P., *et al*. (2003) Combining NLP and probabilistic categorisation for document and term selection for Swiss-Prot medical annotation. *Bioinformatics*, **19** (Suppl. 1), 91–94.

Duda,P., *et al*. (2001) *Pattern Classification*. Wiley, New York.

Dudoit,S., *et al*. (2002) Comparison of discriminant methods for the classification of tumors using gene expression data. *J. Am. Stat. Assoc.*, **97**, 77–87.

Dudoit,S., *et al*. (2003) Multiple hypothesis testing in microarray experiments. *Stat. Sci.*, **18**, 7–103.

Efron,B., *et al*. (2001) Empirical Bayes analysis of a microarray experiment. *J. Am. Stat. Assoc.*, **96**, 1151–1160.

Eom,J. and Zhang,B. (2000) PubMiner:machine learning-based text mining for biomedical information analysis. In *Lecture Notes in Artificial Intelligence*, Vol. 3192, pp. 216–225.

Ferri,F., *et al*. (1994) *Pattern Recognition in Practice IV, Multiple Paradigms, Comparative Studies and Hybrid Systems*, Chapter Comparative Study of Techniques for Large-scale Feature Selection. Elsevier, Amsterdam, pp. 403–413.

Forman,G. (2003) An extensive empirical study of feature selection metrics for text classification. *J. Mach. Learn. Res.*, **3**, 1289–1305.

Fox,R. and Dimmic,M. (2006) A two-sample Bayesian t-test for microarray data. *BMC Bioinformatics*, **7**, 126.

Gabriel,S., *et al*. (2002) The structure of haplotype blocks in the human genome. *Science*, **296**, 2225–2229.

Geurts,P., *et al*. (2005) Proteomic mass spectra classification using decision tree based ensemble methods. *Bioinformatics*, **21**, 3138–3145.

Gevaert,O., *et al*. (2006) Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks. *Bioinformatics*, **22**, e184–e190.

Ghosh,D. and Chinnaiyan,M. (2005) Classification and selection of biomarkers in genomic data using LASSO. *J. Biomed. Biotechnol.*, **2005**, 147–154.

Golub,T., *et al*. (1999)Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Gong,B., *et al*. (2005) Application of genetic algorithm—support vector machine hybrid for prediction of clinical phenotypes based on geneome-wide SNP profiles of sib pairs. In *Lecture Notes in Computer Science 3614*, Springer, Berlin/Heidelberg, pp. 830–835.

Gould,J., *et al*. (2006) Comparative gene marker selection suite. *Bioinformatics*, **22**, 1924–1925.

Guyon,I. and Elisseeff,A. (2003) An introduction to variable and feature selection. *J. Mach Learn Res.*, **3**, 1157–1182.

Guyon,I., *et al*. (2002) Gene selection for cancer classification using support vector machines. *Mach. Learn.*, **46** 389–422.

Hall,M. (1999) Correlation-based feature selection for machine learning. *PhD Thesis.*, Department of Computer Science, Waikato University, New Zealand.

Halperin,E., *et al*. (2005) Tag SNP selection in genotype data for maximizing SNP prediction accuracy. *Bioinformatics*, **21**, i195–i203.

Han,B., *et al*. (2006) Substring selection for biomedical document classification. *Bioinformatics*, **22**, 2136–2142.

He,J. and Zelikovsky,A. (2006) MLR-tagging: informative SNP selection for unphased genotypes based on multiple linear regression. *Bioinformatics*, **22**, 2558–2561.

Hilario,M., *et al*. (2006) Processing and classification of protein mass spectra. *Mass Spectrom. Rev.*, **25**, 409–449.

Holland,J. (1975) *Adaptation in Natural and Artificial Systems*. University of Michigan Press, Ann Arbor.

Inza,I., *et al*. (2000) Feature subset selection by Bayesian networks based optimization. *Artif. Intell.*, **123**, 157–184.

Inza,I., *et al*. (2004) Filter versus wrapper gene selection approaches in DNA microarray domains. *Artif. Intell. Med.*, **31**, 91–103.

Jafari,P. and Azuaje,F. (2006) An assessment of recently published gene expression data analyses: reporting experimental design and statistical factors. *BMC Med. Inform. Decis. Mak.*, **6**, 27.

Jensen,L., *et al*. (2006) Literature mining for the biologist: from information retrieval to biological discovery. *Nat. Rev. Genet.*, **7**, 119–129.

Jiang,H., *et al*. (2004) Joint analysis of two microarray gene-expression data sets to select lung adenocarcinoma marker genes. *BMC Bioinformatics*, **5**, 81.

Jirapech-Umpai,T. and Aitken,S. (2005) Feature selection and classification for microarray data analysis: evolutionary methods for identifying predictive genes. *BMC Bioinformatics*, **6**, 148.

Jong,K., *et al*. (2004) Feature selection in proteomic pattern data with support vector machines. In *Proceedings of the IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, pp. 41–48.

Keles,S., *et al*. (2002) Identification of regulatory elements using a feature selection method. *Bioinformatics*, **18**, 1167–1175.

Kim,S., *et al*. (2006) miTarget: microRNA target gene prediction using a support vector machine. *BMC Bioinformatics*, **7**, 411.

Kittler,J. (1978) *Pattern Recognition and Signal Processing*, *Chapter Feature Set Search Algorithms* Sijthoff and Noordhoff, Alphen aan den Rijn, Netherlands, pp. 41–60.

Kohavi,R., *et al*. (1996) Data mining using MLC++: a machine learning library in C++. In *Tools with Artificial Intelligence*, IEEE Computer Society Press, Washington, DC, pp. 234–245.

Koller,D. and Sahami,M. (1996) Toward optimal feature selection. In *Proceedings of the Thirteenth International Conference on Machine Learning*, Bari, Italy, pp. 284–292.

Kruglyak,L. and Nickerson,D.A. (2001)Variation in the spice of life. *Nat. Genet.*, **27**, 234–236.

LeeP.H., and Shatkay,H. (2006) BNTagger: improved tagging SNP selection using Bayesian networks. *Bioinformatics*, **22**, e211–e219.

Lee,S. and Kang,C. (2004) CHOISS for selection on single nucleotide polymorphism markers on interval regularity. *Bioinformatics*, **20**, 581–582.

Lee,J., *et al*. (2005) An extensive comparison of recent classification tools applied to microarray data. *Comput. Stat. and Data Anal.*, **48**, 869–885.

Lee,K., *et al*. (2003) Gene selection: a Bayesian variable selection approach. *Bioinformatics*, **19**, 90–97.

Leek,J. *et al*. (2006) EDGE: extraction and analysis of differential gene expression. *Bioinformatics*, **22**, 507–508.

Levner,I. (2005) Feature selection and nearest centroid classification for protein mass spectrometry. *BMC Bioinformatics*, **6**, 68.

Li,L. *et al*. (2001) Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method. *Bioinformatics*, **17**, 1131–1142.

Li,L. *et al*. (2004) Applications of the GA/KNN method to SELDI proteomics data. *Bioinformatics*, **20**, 1638–1640.

Li,T., *et al*. (2004) A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. *Bioinformatics*, **20**, 2429–2437.

Li,W. and Yang,Y. (2002) How many genes are needed for a discriminant microarray data analysis? In Lin,S.M. and Johnson,K.F. (eds.), Methods of Microarray Data Analysis. *First Conference on Critical Assessment of Microarray Data Analysis, CAMDA2000*, pp. 137–150.

Li,X., *et al*. (2005) Large-scale ensemble decision analysis of sib-pair ibd profiles for identification of the relevant molecular signatures for alcoholism. In *Lecture Notes in Computer Science 3614*, Springer, Berlin/Heidelberg, pp. 1184–1189.

Lin,Z. and Altman,R.B. (2004) Finding haplotype tagging SNPs by use of principal components analysis. *Am. J. Hum. Genet.*, **73**, 850–861.

Liu,H. and Motoda,H. (1998) *Feature Selection for Knowledge Discovery and Data Mining*. Kluwer Academic Publishers, Norwell, MA.

Liu,H., *et al*. (2002) A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns. *Genome Inform.*, **13**, 51–60.

Liu,H., *et al*. (2004) Using amino acid patterns to accurately predict translation initiation sites. In *Silico Biol.*, **4**, 255–269.

Lyons-Weiler,J. *et al*. (2004) Tests for finding complex patterns of differential expression in cancers: towards individualized medicine. *BMC Bioinformatics*, **5**, 110.

Downloaded from http://bioinformatics.oxfordjournals.org/ at milner library illinois state university on December 4, 2012

Mamitsuka,H. (2006) Selecting features in microarray classification using ROC curves. *Pattern Recognit.*, **39**, 2393–2404.

Ma,S. and Huang,J. (2005) Regularized ROC method for disease classification and biomarker selection with microarray data. *Bioinformatics*, **21**, 4356–4362.

Medina,I., *et al.* (2007) Prophet, a web-based tool for class prediction using microarray data. *Bioinformatics*, **23**, 390–391.

Molinaro,A., *et al.* (2005) Prediction error estimation: a comparison of resampling methods. *Bioinformatics*, **21**, 3301–3307.

Newton,M., *et al.* (2001) On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *J. Comput. Biol.*, **8**, 37–52.

Ooi,C. and Tan,P. (2003) Genetic algorithms applied to multi-class prediction for the analysis of gene expression data. *Bioinformatics*, **19**, 37–44.

Pan,W. (2003) On the use of permutation in and the performance of a class of nonparametric methods to detect differential gene expression. *Bioinformatics*, **19**, 1333–1340.

Park,P., *et al.* (2001) A nonparametric scoring algorithm for identifying informative genes from microarray data. *Pac. Symp. on Biocompu.*, **6**, 52–63.

Pavlidis,P. and Poirazi,P. (2006) Individualized markers optimize class prediction of microarray data. *BMC Bioinformatics*, **7**, 345.

Petricoin,E. and Liotta,L. (2003) Mass spectometry-based diagnostic: the upcoming revolution in disease detection. *Clin. Chem.*, **49**, 533–534.

Petricoin,E., *et al.* (2002) Use of proteomics patterns in serum to identify ovarian cancer. *The Lancet*, **359**, 572–577.

Ploner,A., *et al.* (2006) Multidimensional local false discovery rate for microarray studies. *Bioinformatics*, **22**, 556–565.

Pounds,S. and Cheng,C. (2004) Improving false discovery rate estimation. *Bioinformatics*, **20**, 1737–1754.

Prados,J., *et al.* (2004) Mining mass-spectra for diagnosis and biomarker discovery of cerebral accidents. *Proteomics*, **4**, 2320–2332.

Ressom,H., *et al.* (2005) Analysis of mass spectral serum profiles for biomarker selection. *Bioinformatics*, **21**, 4039–4045.

Ressom,H., *et al.* (2007) Peak selection from MALDI-TOF mass spectra using ant colony optimization. *Bioinformatics*, **23**, 619–626.

Ross,D., *et al.* (2000) Systematic variation in gene expression patterns in human cancer cell lines. *Nat. Genet.*, **24**, 227–234.

Ruiz,R., *et al.* (2006) Incremental wrapper-based gene selection from microarray data for cancer classification. *Pattern Recognit.*, **39**, 2383–2392.

Saeys,Y., *et al.* (2004) Feature selection for splice site prediction: a new method using EDA-based feature ranking. *BMC Bioinformatics*, **5**, 64.

Saeys,Y., *et al.* (2007) In search of the small ones: improved prediction of short exons in vertebrates, plants, fungi, and protists. *Bioinformatics*, **23**, 414–420.

Salzberg,S., *et al.* (1998) Microbial gene identification using interpolated markov models. *Nucleic Acids Res.*, **26**, 544–548.

Scheid,S. and Spang,R. (2005) twilight; a Bioconductor package for estimating the local false discovery rate. *Bioinformatics*, **21**, 2921–2922.

Shah,S. and Kusiak,A. (2004) Data mining and genetic algorithm based gene/SNP selection. *Artif. Intell. Med.*, **31**, 183–196.

Sham,P., *et al.* (2007) Combining functional and linkage disequilibrium information in the selection of tag snps. *Bioinformatics*, **23**, 129–131.

Shin,H. and Markey,M. (2006) A machine learning perspective on the development of clinical decision support systems utilizing mass spectra of blood samples. *J. Biomed. Inform.*, **39**, 227–248.

Siedelecky,W. and Sklansky,J. (1998) On automatic feature selection. *Int. J. Pattern Recogni.*, **2**, 197–220.

Sima,C. and Dougherty,E. (2006) What should be expected from feature selection in small-sample settings. *Bioinformatics*, **22**, 2430–2436.

Sima,C., *et al.* (2005) Superior feature-set ranking for small samples using bolstered error estimation. *Bioinformatics*, **21**, 1046–1054.

Sinha,S. (2003) Discriminative motifs. *J. Comput. Biol.*, **10**, 599–615.

Skalak,D. (1994) Prototype and feature selection by sampling and random mutation hill climbing algorithms. In *Proceedings of the Eleventh International Conference on Machine Learning*, pp. 293–301.

Smyth,G. (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. in Genet. and Mol. Biol.*, **3**, Article 3.

Somorjai,R., *et al.* (2003) Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions. *Bioinformatics*, **19**, 1484–1491.

Statnikov,A., *et al.* (2005) A comprhensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. *Bioinformatics*, **21**, 631–643.

Storey,J. (2002) A direct approach to false discovery rates. *J. R. Stat. Soc. Ser. B*, **64**, 479–498.

Su,Y., *et al.* (2003) RankGene: identification of diagnostic genes based on expression data. *Bioinformatics*, **19**, 1587–1579.

Tadesse,M., *et al.* (2004) Identification of DNA regulatory motifs using Bayesian variable selection. *Bioinformatics*, **20**, 2553–2561.

Thomas,J., *et al.* (2001) An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles. *Genome Res.*, **11**, 1227–1236.

Tibshirani,R., *et al.* (2004) Sample classification from protein mass spectrometry, by 'peak probability contrast'. *Bioinformatics*, **20**, 3034–3044.

Trevino,V. and Falciani,F. (2006) GALGO: an R package for multivariate variable selection using genetic algorithms. *Bioinformatics*, **22**, 1154–1156.

Troyanskaya,O., *et al.* (2002) Nonparametric methods for identifying differentially expressed genes in microarray data. *Bioinformatics*, **18**, 1454–1461.

Tusher,V., *et al.* (2001) Significance analysis of microarrays applied to ionizing radiation response. In *Proceedings of the National Academy of Sciences*, Vol. 98, pp. 5116–5121.

Varshavsky,R., *et al.* (2006) Novel unsupervised feature filtering of biological data. *Bioinformatics*, **22**, e507–e513.

Wang,Y., *et al.* (2005) Gene selection from microarray data for cancer classification–a machine learning approach. *Comput. Biol. Chem.*, **29**, 37–46.

Wang,Y., *et al.* (2006) Tumor classification based on DNA copy number aberrations determined using SNPS arrays. *Oncol. Rep.*, **5**, 1057–1059.

Weston,J., *et al.* (2003) Use of the zero-norm with linear models and kernel methods. *J. Mach. Learn. Res.*, **3**, 1439–1461.

Witten,I. and Frank,E. (2005) *Data Mining: Practical Machine Learning Tools and Techniques,* 2nd Ed. Morgan Kaufmann, San Francisco.

Wu,B., *et al.* (2003) Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. *Bioinformatics*, **19**, 1636–1643.

Xing,E.P. *et al.* (2001) Feature selection for high-dimensional genomic microarray data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pp. 601–608.

Xiong,M., *et al.* (2001) Biomarker identification by feature wrappers. *Genome Res.*, **11**, 1878–1887.

Yang,Y., *et al.* (2005) Identifying differentially expressed genes from microarray experiments via statistic synthesis. *Bioinformatics*, **21**, 1084–1093.

Yeoh,E., *et al.* (2002) Classification, subtype discovery, and prediction of outcome in pediatric lymphoblastic leukemia by gene expression profiling. *Cancer Cell*, **1**, 133–143.

Yeung,K. and Bumgarner,R. (2003) Multiclass classification of microarray data with repeated measurements: application to cancer. *Genome Biol.*, **4**, R83.

Yeung,K., *et al.* (2005) Bayesian model averaging: development of an improved multi-class, gene selection and classification tool for microarray data. *Bioinformatics*, **21**, 2394–2402.

Yu,J. and Chen,X. (2005) Bayesian neural network approaches to ovarian cancer identification from high-resolution mass spectrometry data. *Bioinformatics*, **21**, i487–i494.

Yu,J., *et al.* (2005) Ovarian cancer identification based on dimensionality reduction for high-throughput mass spectrometry data. *Bioinformatics*, **21**, 2200–2209.

Yu,L. and Liu,H. (2004) Efficient feature selection via analysis of relevance and redundancy. *J. Mach. Learn. Res.*, **5**, 1205–1224.

Zavaljevsky,N., *et al.* (2002) Support vector machines with selective kernel scaling for protein classification and identification of key amino acid positions. *Bioinformatics*, **18**, 689–696.

Zhang,X., *et al.* (2006) Recursive SVM feature selection and sample classification for mass-spectrometry and microarray data. *BMC Bioinformatics*, **7**, 197.

Downloaded from http://bioinformatics.oxfordjournals.org/ at milner library illinois state university on December 4, 2012