Netw Model Anal Health Inform Bioinforma (2012) 1:47–61
DOI 10.1007/s13721-012-0006-6

ORIGINAL ARTICLE

Threshold-based feature selection techniques for high-dimensional bioinformatics data

Jason Van Hulse · Taghi M. Khoshgoftaar · Amri Napolitano · Randall Wald

Received: 16 November 2011 / Revised: 12 March 2012 / Accepted: 12 March 2012 / Published online: 22 May 2012

Abstract Analysis conducted for bioinformatics applications often requires the use of feature selection methodologies to handle datasets with very high dimensionality. We propose 11 new threshold-based feature selection techniques and compare the performance of these new techniques to that of six standard filter-based feature selection procedures. Unlike other comparisons of feature selection techniques, we directly compare the feature rankings produced by each technique using Kendall's Tau rank correlation, showing that the newly proposed techniques exhibit substantially different behaviors than the standard filter-based feature selection methods. Our experiments consider 17 different bioinformatics datasets, and the similarities of the feature selection techniques are analyzed using the Frobenius norm. The feature selection techniques are also compared by using Naive Bayes and Support Vector Machine algorithms to learn from the training datasets. The experimental results show that the new procedures perform very well compared to the standard filters, and hence are useful feature selection methodologies for the analysis of bioinformatics data.

J. Van Hulse · T. M. Khoshgoftaar (✉) · A. Napolitano · R. Wald
Data Mining and Machine Learning Laboratory, Department of Computer and Electrical Engineering and Computer Science, Florida Atlantic University, Boca Raton, FL 33431, USA
e-mail: khoshgof@fau.edu

J. Van Hulse
e-mail: jvanhulse@gmail.com

A. Napolitano
e-mail: amrifau@gmail.com

R. Wald
e-mail: rdwald@gmail.com

Keywords Threshold-based feature selection · Bioinformatics · Frobenius norm · Kendall's Tau rank correlation · Correlation matrix · Feature selection

1 Introduction

Data mining and machine learning techniques have become increasingly important in the field of bioinformatics as the amount of data collected and stored has grown dramatically. As in other application domains that rely on data mining techniques, there is often too much data for human analysis—hundreds if not thousands of features are being examined, and it is often very difficult to draw conclusions through direct examination of the data. A variety of biological data sources are amenable to bioinformatic analysis; for example, protein sequences (Radivojac et al. 2004) and properties (Lee et al. 2007; Gupta et al. 2008), genetic codes (Sun et al. 2006; Akbani 2005) and mass spectroscopy results (Li et al. 2004). Of particular interest is the problem of analyzing microarray data. Each microarray consists of a so-called "gene chip" with thousands of probes for different genes; a DNA or RNA sample is placed on the chip, and each gene glows corresponding to how active that gene was in the sample (Piatetsky-Shapiro 2003).

One common and pernicious problem often encountered in bioinformatics datasets is the overabundance of features or attributes. *Feature selection* methods are widely used to reduce the number of features. Given a dataset D with a set of features $\mathbb{F} = (X^1, \ldots, X^m)$, the objective of feature selection is to select a subset of features $\mathbb{F}^0 = (X^{j_1}, \ldots, X^{j_p})$ such that $p \ll m$ and $\mathbb{F}^0$ satisfies the particular conditions of the task at hand. In a classification problem, for example, the objective may be to extract the set of features that

maximize the classification accuracy of the learner. The benefits of feature selection include faster model training, reduced susceptibility to overfitting, offsetting the harmful effects of the curse of dimensionality, and reducing storage, memory, and processing requirements during data analysis (Guyon and Elisseeff 2003). The obvious drawback of these techniques, however, is the possibility that critical attributes may be omitted, thereby hurting performance.

We propose a set of threshold-based feature selection (TBFS) techniques which substantially extend the FAST algorithm proposed by Chen and Wasikowski (2008). FAST is based on the area under a ROC curve (AUC) generated by moving the decision boundary of a single feature classifier with thresholds placed using an even-bin distribution. More specifically, each attribute is partitioned into bins with an equal number of examples, and within each bin, the true positive and false positive rates are calculated. The area under the ROC curve is then computed across all bins. TBFS pairs each attribute with the class, normalizes the attribute to have a range between 0 and 1, and classifies examples as either positive or negative at each possible threshold. A variety of metrics ω, which provide the relative importance of the attribute vis-à-vis the class, are then calculated. TBFS is much more general than the procedure proposed by Chen and Wasikowski. FAST calculates a ROC curve by binning the attribute, while TBFS does not require discretization, making it more precise and eliminating the often vexing question of how wide the bins should be. Further, there are 11 different versions of TBFS which are based on 11 different metrics for feature ranking. Another useful property of TBFS is that it can easily be extended to incorporate additional metrics.

Feature selection methods are often considered in the context of classification, but they are useful in other situations as well. In gene expression array studies, for example, biologists might apply feature selection techniques to identify which genes are most important. Therefore, in addition to considering the resulting classification performance of the different filter-based feature selection techniques we analyze, this study also directly compares the attribute rankings produced by these methods using the Kendall's Tau rank correlation statistic (see Sect. 6). By measuring the rank correlation between the attribute rankings, it is easier to discern which techniques produce similar results regardless of the ultimate use of the data. For example, feature selection techniques A and B may result in similar accuracy when used in conjunction with classifier Z, but will the same results hold if classifiers X or Y are used? In such a situation, it is not clear if feature selection techniques A and B are truly similar for all classifiers or only for classifier Z. In order to better understand the true similarities among feature selection techniques, the attribute rankings are directly compared to one another. While other studies have considered the feature rankings produced by feature selection techniques (see Sect. 2 for further discussion), to the best of our knowledge, our study is unique in this respect.

In summary, the main contribution of this work, which is a substantial expansion of the work published in IRI 2011 (Van Hulse et al. 2011), is as follows. First, the 11 versions of TBFS are described in detail. The correlation in attribute rankings produced by these 11 versions are analyzed and compared to those produced by six commonly used filter-based feature selection techniques using 17 different high-dimensional bioinformatics datasets. All 17 feature selection methods (Table 1) are also evaluated using two different learning algorithms, Naive Bayes and Support Vector Machines.

The remainder of this work is organized as follows. Section 2 describes related work, while six common filter-based feature selection techniques are presented in Sect. 3. Our proposed TBFS methodology is presented in Sect. 4. The datasets used in the experiments are described in Sect. 5, while Kendall's Tau rank correlation is explained in Sect. 6. Experimental results are provided in Sect. 7 followed by conclusions and directions for future work in Sect. 8.

2 Related work

Much of the research on microarray analysis has focused on improving classification models used to categorize unknown samples as "healthy" or "sick." Since the data have literally thousands of features, with commonly-used sets having 2,000 (Alon et al. 1999) and 7,129 (Golub et al. 1999) features each, feature selection is a necessity. A variety of techniques have been employed for the selection step. Some previous researchers utilize the standard array of feature ranking and subset evaluation filters and wrappers, coupled with traditional data mining techniques; these either analyze filters alone (Furey et al. 2000; Zhang et al. 2001; Xing et al. 2001; Kamal et al. 2009) or compare filters and wrappers (Inza et al. 2004; Wang et al. 2005) or filters, wrappers and principal component analysis (Model 2001). Others employ genetic algorithms (Li et al. 2001; Jirapech-Umpai 2005) or minimum redundancy (Ding and Peng 2003; Yu 2004; Peng et al. 2005) to find the optimal subset of genes for classification purposes. And some propose novel feature selection techniques specifically designed for microarray analysis (Peddada et al. 2003; Wang et al. 2005).

One common feature of previous research is that the quality of the feature selection techniques is evaluated by testing their ability to classify the data. The feature

Table 1 List of 17 Filter-based feature selection techniques (including 11 new threshold-based procedures)

Abbreviation	Name
Standard filter-based feature selection techniques	
χ^2	χ^2 statistic
GR	Gain ratio
IG	Information gain
RF	ReliefF
RFW	ReliefF-weight by distance
SU	Symmetric uncertainty
Threshold-based feature selection techniques	
F	F-measure
OR	Odds ratio
Pow	Power
PR	Probability ratio
GI	Gini Index
MI	Mutual information
KS	Kolmogorov–Smirnov statistic
Dev	Deviance
GM	Geometric mean
AUC	Area under the ROC curve
PRC	Area under the precision-recall curve

rankings themselves might undergo some ad hoc comparisons (Wang et al. 2005; Li et al. 2001; Peddada et al. 2003; Jirapech-Umpai 2005), but no systemic study comparing feature selection algorithms on their own has been conducted. The closest the literature comes to this approach is in evaluating feature clustering techniques (Datta 2003; Brown et al. 2000; Au et al. 2005); these also treat genes as features while focusing specifically on finding novel genes which are especially relevant to the underlying biological problem. Kalousis et al. (2007) measure the stability and robustness of feature selection methods relative to perturbations of the training set. They consider, among other things, Spearman's rank correlation between the feature ranking lists produced by the same learner for different samples of the training dataset. This work differs from ours in that we consider the correlation between two different filters using the same dataset, while (Kalousis et al. 2007) looks at the correlation for a single filter over different samples of the same dataset. Nonetheless, the idea of evaluating the intrinsic feature ranking is the same. Stability of feature selection techniques is also considered in other work (Davis et al. 2006; Křížek et al. 2007; Kunchev 2007). Saeys et al. (2008) utilizes an ensemble of feature selection methods to improve stability of the selected attribute set.

Previous work (Van Hulse et al. 2009) presented very preliminary findings using 3 of the 11 techniques discussed in this work. This work is a substantial extension of this previous work since it not only proposes eight additional TBFS techniques, but also dramatically expands the number of datasets in the empirical study from 5 to 17. Further, this previous work was focused on feature selection with imbalanced data, and in accordance with that objective we examined the impact of four different data sampling methods on the proposed techniques. In this work, though some of the datasets are imbalanced (see Table 2), we do not consider data sampling but instead concentrate on exploring the relationships between the different filters.

Table 2 Bioinformatics datasets

Dataset name	Abbreviation	Attributes (#)	Total (#)	Positive (#)	Positive(%)
ECML pancreas	ECML	27,680	90	8	8.9
Central nervous system	CNS	7,130	60	21	35.0
Colon	Colon	2,001	62	22	35.5
DLBCL tumor	Tum	7,130	77	19	24.7
Lymphoma	Lymph	4,027	96	23	24.0
DLBCL	DLB	4,027	47	23	48.9
Lung cancer	LC	12,534	181	31	17.1
Acute lymphoblastic leukemia	ALL	12,559	327	79	24.2
Prostate	Pros	12,601	136	59	43.4
MLL leukemia	MLL	12,583	72	20	27.8
Breast cancer	Brst	24,482	97	46	47.4
All AML leukemia	AAL	7,130	72	25	34.7
Translation initiation	Tran	925	13,375	3,312	24.8
Ovarian cancer	Ov	15,155	253	91	36.0
DLBCL NIH	NIH	7,399	240	103	42.9
Lung	Lung	12,601	203	65	32.0
Brain tumor	Brain	27,679	90	23	25.6

Finally, this previous work did not utilize classifiers to compare the performances of the different filters.

3 Standard filter-based feature selection techniques

The standard filter-based feature ranking techniques considered in this work include chi-squared (Witten and Frank 2005), information gain (Hall and Holmes 2003; Quinla 1993; Witten and Frank 2005), gain ratio (Quinla 1993; Witten 2005), two versions of ReliefF (Kononenk 1994; Kira and Rendell 1992), and symmetric uncertainty (Hall and Holmes 1999; Witten 2005). All of these feature selection methods are available within Weka (Witten 2005). Since these methods are widely known, we provide only a brief summary; the interested reader should consult with the included references for further details.

The Chi-squared method (χ^2) utilizes the χ^2 statistic to measure the strength of the relationship between each independent variable and the class. Information gain (IG) determines the significance of a feature based on the amount by which the entropy of the class decreases when considering that feature. Gain Ratio (GR) is a refinement of IG, adjusting for features that have a large number of values. GR attempts to maximize the information gain of the feature while minimizing the number of values. Symmetric uncertainty (SU) also adjusts IG to account for attributes with more values, and normalizes its value to lie in the range [0, 1]. These techniques utilize the method of Fayyad and Irani (1992) to discretize continuous attributes, and all four methods are bivariate, considering the relationship between each attribute and the class, excluding the other independent variables.

Relief randomly samples an example from the data and finds 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 relevance scores for each attribute. This process is repeated for m examples, as specified by the user. ReliefF (RF) extends Relief by handling noise and multiclass data sets (Kononenk 1994). RF is implemented within Weka (Witten 2005) with the "weight nearest neighbors by their distance" parameter set to false. ReliefF-W (RFW) is similar to RF except the "weight nearest neighbors by their distance" parameter is set to true.

4 Threshold-based feature selection techniques

This section describes the TBFS method for feature ranking. Similar to the χ^2, IG, GR and SU, TBFS is a bivariate procedure; each attribute is evaluated against the class, independent of all other features in the dataset. After normalizing each attribute to have a range between 0 and 1, simple classifiers are built for each threshold value $t \in [0, 1]$ according to two different classification rules. For classification rule 1, examples with a normalized value greater than t are classified P while examples with a normalized value less than t are classified as N (assuming each instance x is assigned to one of two classes $c(x) \in \{P, N\}$). For classification rule 2, examples with a normalized value greater than t are classified N while examples with a normalized value less than t are classified as P. Two different classification rules must be considered to account for the fact that for some attributes, large values of the attribute may have a greater correlation with the positive class, while for other attributes, large values of the attribute may have a greater correlation with the negative class. Metric ω is calculated using the formulas provided in Algorithm 1 either at each threshold t or across all thresholds for both classification rules. Finally, the metric resulting from the classification rule which provides the largest value is used as the relevancy measure for that attribute relative to metric ω.

Many of the metrics ω (e.g., AUC, PRC, GM, F, KS) are primarily used to measure the performance of classification models, using the posterior probabilities computed by such models to classify examples as either negative or positive depending on the classification threshold. The normalized attribute values can be thought of as posterior probabilities, e.g., $p(P \mid x) = \hat{X}^j(x)$ for classification rule 1, and the metrics ω are computed against this "posterior". Intuitively, attributes where positive and negative examples are evenly distributed along the distribution of X produce weak measures ω and poor relevancy scores in a similar manner that poor predictive models have positive and negative examples evenly distributed along the distribution of the posterior probability produced by the model. Note further that TBFS can easily be extended to include additional metrics. As mentioned previously, TBFS is a substantial extension of the FAST algorithm (Chen and Wasikowski 2008). FAST only utilizes the area under the ROC curve, and Chen and Wasikowski discretize X^j when computing the AUC.

For additional information on the AUC, PRC, KS, GM, and F metrics, see Seliya et al. (2009), Witten (2005), Conove (1971). The F-measure uses the tunable parameter β to weight precision (PRE) and recall or true positive rate (TPR). In this work, $\beta = 1$ is used, and hence F is the harmonic mean of recall and precision. The Gini index was introduced by Breiman et al. (1984). Pow, OR and PR were used by Forman (2003) in the context of feature selection for text categorization. Additional information on deviance can be found in Khoshgoftaar et al. (2002) and MI is

Algorithm 1: Threshold-Based Feature Selection Algorithm

input :

a. Dataset D with features $X^j, j = 1, \ldots, m$;

b. Each instance $x \in D$ is assigned to one of two classes $c(x) \in \{P, N\}$;

c. $|P| = |\{x \in D | c(x) = P\}|$, $|N| = |\{x \in D | c(x) = N\}|$;

d. The value of attribute X^j for instance x is denoted $X^j(x)$;

e. Metric $\omega \in$ {F, OR, Pow, PR, GI, MI, KS, Dev, GM, AUC, PRC}.

output: Ranking $\mathbb{R} = \{r^1, r^2, \ldots, r^m\}$ where attribute X^j is the r^j-th most significant attribute as determined by metric ω.

for $X^j, j = 1, \ldots, m$ **do**

Normalize $X^j \mapsto \hat{X}^j = \frac{X^j - \min(X^j)}{\max(X^j) - \min(X^j)}$, $\hat{X}^j \in [0, 1]$;

for $t \in [0, 1]$ **do**

Compute Basic Metrics:

Classification Rule 1: $\forall\, x \in D, \hat{c}^t(x) = P \iff \hat{X}^j(x) > t$, otherwise $\hat{c}^t(x) = N$.

$TP(t) = \left|\{x | (\hat{c}^t(x) = P) \cap (c(x) = P)\}\right|$, $TN(t) = \left|\{x | (\hat{c}^t(x) = N) \cap (c(x) = N)\}\right|$,

$FP(t) = \left|\{x | (\hat{c}^t(x) = P) \cap (c(x) = N)\}\right|$, $FN(t) = \left|\{x | (\hat{c}^t(x) = N) \cap (c(x) = P)\}\right|$,

$TPR(t) = \frac{|TP(t)|}{|P|}$, $TNR(t) = \frac{|TN(t)|}{|N|}$, $FPR(t) = 1 - TNR(t)$, $FNR(t) = 1 - TPR(t)$,

$PRE(t) = \frac{|TP(t)|}{|TP(t)| + |FP(t)|}$, $NPV(t) = \frac{|TN(t)|}{|TN(t)| + |FN(t)|}$.

Compute Final Metrics:

Metric ω		Calculation
$\mathrm{F}^1(\hat{X}^j)$	=	$\max_{t \in [0,1]} \frac{(1+\beta^2) PRE(t) TPR(t)}{\beta^2 PRE(t) + TPR(t)}$; $\beta = 1$
$\mathrm{OR}^1(\hat{X}^j)$	=	$\max_{t \in [0,1]} \frac{TP(t) * TN(t)}{FP(t) * FN(t)}$
$\mathrm{Pow}^1(\hat{X}^j)$	=	$\max_{t \in [0,1]} ((1 - FPR(t))^k - (1 - TPR(t))^k)$; $k = 5$
$\mathrm{PR}^1(\hat{X}^j)$	=	$\max_{t \in [0,1]} \frac{TPR(t)}{FPR(t)}$
$\mathrm{GI}^1(\hat{X}^j)$	=	$\min_{t \in [0,1]} [2PRE(t)(1 - PRE(t)) + 2NPV(t)(1 - NPV(t))]$
$\mathrm{KS}^1(\hat{X}^j)$	=	$\max_{t \in [0,1]} \lvert TPR(t) - FPR(t) \rvert$
$\mathrm{GM}^1(\hat{X}^j)$	=	$\max_{t \in [0,1]} \sqrt{TPR(t) \times TNR(t)}$
$\mathrm{AUC}^1(\hat{X}^j)$	=	Area under the curve generated by $(FPR(t), TPR(t)), t \in [0, 1]$
$\mathrm{PRC}^1(\hat{X}^j)$	=	Area under the curve generated by $(PRE(t), TPR(t)), t \in [0, 1]$
$\mathrm{MI}^1(\hat{X}^j)$	=	$\max_{t \in [0,1]} \sum_{\hat{c}^t \in \{P,N\}} \sum_{c \in \{P,N\}} p(\hat{c}^t, c) \log \frac{p(\hat{c}^t, c)}{p(\hat{c}^t) p(c)}$ where $p(\hat{c}^t = \alpha, c = \beta) = \frac{\lvert\{x \lvert (\hat{c}^t(x) = \alpha) \cap (c(x) = \beta)\}\rvert}{|P| + |N|}$, $p(\hat{c}^t = \alpha) = \frac{\lvert\{x \lvert \hat{c}^t(x) = \alpha\}\rvert}{|P| + |N|}$, $p(c = \alpha) = \frac{\lvert\{x \lvert c(x) = \alpha\}\rvert}{|P| + |N|}$, $\alpha, \beta \in \{P, N\}$
$\mathrm{Dev}^1(\hat{X}^j)$	=	$\min_{t \in [0,1]} \left[\sum_{x \in S^t} (v(x) - v(S^t))^2 + \sum_{x \in \bar{S}^t} (v(x) - v(\bar{S}^t))^2\right]$ where $S^t = \{x \lvert \hat{X}^j > t\}$, $\bar{S}^t = \{x \lvert \hat{X}^j \leq t\}$, $v(S^t) = \lvert S^t \rvert^{-1} \sum_{x \in S^t} v(x)$, $v(\bar{S}^t) = \lvert \bar{S}^t \rvert^{-1} \sum_{x \in \bar{S}^t} v(x)$, and $v(x) = 1$ if $x \in P$, otherwise, $v(x) = 0$

Compute the same basic metrics and final metrics (denoted as ω^2) as listed above, but using:

Classification Rule 2: $\forall\, x \in D, \hat{c}^t(x) = N \iff \hat{X}^j(x) > t$, otherwise $\hat{c}^t(x) = P$.

$\omega(\hat{X}^j) = \max\left(\omega^1(\hat{x}^j), \omega^2(\hat{x}^j)\right)$

Create attribute ranking $\mathbb{R}$ using $\omega(\hat{X}^j) \forall j$

Fig. 1 Algorithm 1: Threshold-based feature selection algorithm

utilized in Battiti (1994). Additional derivations for the KS metric, which is relatively less common than some of the other metrics, and GI, whose formula in Algorithm 1 underwent a substantial transformation from what is normally described, are provided in the appendix (Sect. 8).

5 Datasets

The datasets utilized in our experiments are listed in Table 2. All of the datasets come from the bioinformatics application domain, and all but two (Translation and Ovarian) are microarray expression datasets. Table 2 provides the number of attributes, number of total examples, number of positive examples and the percentage of positive examples for each dataset. Note that all of the datasets used in this work have a binary dependent variable. Further note that these datasets exhibit a wide distribution of class skew (i.e., the percentage of positive examples).

For the ovarian cancer dataset (Petricoin III et al. 2002), the researchers took serum samples from patients with and without cancer and ran them through a mass spectroscopy machine, giving them 15155 separate mass/charge values. That is, all the proteins in the sample were ionized and deflected through a magnetic field such that proteins with a

different ratio of mass to charge would behave differently and thus be detected separately. Thus, the different mass/charge values reflect the relevant abundance of different proteins in each serum sample.

The Translation dataset (Pedersen and Nielsen 1997) is based on a set of mRNA sequences for different genes found in vertebrates and plants, each of which is annotated with its translation initiation point (the ATG which represents the start of the gene). To generate features for a given instance, the upstream and downstream parts of the sequence (that is, the parts before and after the translation initiation point) are searched for all 20 amino acids and the stop codon, as well as for all two-amino-acid sequences (and for the pairs which include the stop codon). The number of times each amino acid or pair of amino acids is found is the value for that feature.

6 Kendall's Tau rank correlation

Kendall's Tau rank correlation statistic (Conove 1971) is used to measure the degree of similarity between the attribute rankings of two techniques. Suppose the rankings r_1 and r_2 are being compared. For each attribute j in the dataset there is an ordered pair $(r_1(j), r_2(j))$ where $r_1(j)$ and $r_2(j)$ are the rankings of attribute j produced by r_1 and r_2. For each pair of attributes (j_1, j_2) the rankings $(r_1(j_1), r_2(j_1))$ and $(r_1(j_2), r_2(j_2))$ are compared and given a value of $+1$ or -1 depending on whether the two rankings are concordant or discordant. Assuming that both r_1 and r_2 do not contain tied ranks, a pair of attributes (j_1, j_2) is considered *concordant* if $r_1(j_1) > r_1(j_2)$ and $r_2(j_1) > r_2(j_2)$ or $r_1(j_1) < r_1(j_2)$ and $r_2(j_1) < r_2(j_2)$. Otherwise, (j_1, j_2) are said to be *discordant*. There are a total of $\frac{n(n-1)}{2}$ pairs of attributes. If S is the sum of the scores for each pair of attributes as determined by their concordance or discordance, Kendall's Tau is calculated as $\tau = S/\frac{n(n-1)}{2}$. If all pairs are concordant, then the two rankings are in complete agreement and $\tau = 1$. If the two rankings are exactly opposite, (i.e., $\forall j, r_2(j) = n - r_1(j) + 1$), then all pairs will be discordant and $\tau = -1$. If τ is close to zero, then the correlation between the two rankings is very weak.

7 Empirical results

7.1 Correlation among feature ranking methods

Figure 1 presents the correlation matrix for the 17 feature selection techniques using Kendall's Tau rank correlation. The entries above the diagonal represent τ averaged over all 17 datasets, while entries below the diagonal represent the standard deviation of τ over the same datasets. For example, the average rank correlation between χ^2 and GR is 0.92 with a standard deviation of 0.10. Though we present some details for a few individual datasets below, it is impractical and redundant to present a separate correlation matrix for each dataset, so we chose to present

Std\Avg	χ2	GR	IG	RF	RFW	SU	F	OR	Pow	PR	GI	MI	KS	Dev	GM	AUC	PRC
χ2		**.92**	**.95**	.15	.14	**.94**	.17	.22	.20	.13	.17	.27	.23	.26	.20	.22	.25
GR	.10		**.93**	.14	.13	**.93**	.16	.22	.19	.13	.17	.24	.21	.24	.17	.20	.22
IG	.06	.09		.15	.14	**.94**	.17	.22	.20	.13	.17	.27	.23	.26	.20	.22	.24
RF	.12	.10	.12		.60	.15	.21	.27	.30	.19	.13	.35	.34	.32	.33	.29	.36
RFW	.11	.09	.11	.13		.14	.20	.22	.22	.13	.10	.28	.28	.25	.27	.24	.28
SU	.09	.11	.10	.11	.11		.17	.23	.20	.13	.17	.26	.22	.25	.19	.21	.24
F	.14	.12	.14	.18	.14	.13		.38	.16	.04	.00	.52	.56	.46	.49	.50	.36
OR	.16	.16	.16	.12	.11	.17	.20		.46	.38	.26	.58	.45	.58	.36	.40	.53
Pow	.21	.18	.20	.10	.09	.20	.23	.17		.52	.31	.48	.39	.51	.36	.40	.64
PR	.12	.12	.11	.11	.08	.12	.21	.21	.16		.64	.25	.20	.32	.19	.20	.40
GI	.14	.15	.14	.07	.06	.15	.17	.13	.12	.27		.16	.12	.24	.10	.12	.29
MI	.23	.19	.24	.13	.12	.23	.22	.10	.17	.14	.09		.73	.75	.59	.60	.65
KS	.20	.16	.21	.13	.13	.19	.25	.14	.16	.14	.07	.09		.65	.77	.72	.64
Dev	.26	.22	.25	.14	.13	.25	.23	.17	.19	.15	.10	.19	.15		.52	.55	.64
GM	.16	.13	.17	.14	.13	.16	.23	.16	.16	.16	.07	.09	.07	.14		.68	.58
AUC	.19	.15	.19	.16	.14	.18	.21	.13	.13	.11	.07	.14	.10	.16	.08		.64
PRC	.22	.18	.21	.11	.11	.21	.24	.16	.13	.16	.08	.13	.13	.24	.12	.11	

Fig. 1 Correlation matrix for 17 filter-based feature selection techniques

summarized information in Fig. 1. If the mean correlation is greater than 0.9, then the cell in Fig. 1 is in bold text. For entries with a mean correlation between 0.5 and 0.9, then cell background is highlighted.

Figure 1 is divided into four quadrants. The top-left quadrant can be used to compare the six standard filter-based feature selection techniques, while the bottom-right quadrant compares the 11 new TBFS techniques to one another. Among the standard filters there are clearly two distinct groups—in the first group, GR, IG, χ^2 and SU are all highly correlated, while in the second group, RF and RFW are moderately correlated. The correlation among the new TBFS techniques is generally moderate or low. PR and GI are the two TBFS filters that are least correlated with the nine other techniques, and therefore these two filters provide the most divergent attribute rankings compared to the other filters. The six TBFS filters GM, AUC, PRC, KS, Dev and MI form a cluster with average correlations between 0.5 and 0.9, though even among these techniques, there is substantial divergence in the attribute rankings with correlations often below 0.75. We conclude that while there is a moderate degree of correlation among some of the 11 new TBFS techniques, each technique provides a different interpretation of the relative value of the features.

The top-right and bottom-left quadrants can be used to compare the six standard filter-based techniques to the 11 new TBFS techniques. The top-right quadrant is the mean correlation over all 17 datasets, while the bottom-left quadrant is the standard deviation of τ over the 17 datasets. It is important to understand if the information provided by the 11 TBFS techniques is different than that provided by the standard techniques—if they are highly correlated, then there is no need to consider 11 new procedures for feature selection. Notably, all 11 TBFS techniques exhibit very low correlation with the standard techniques. Among the 66 pairwise correlations, none were greater than 0.5, and in fact, the highest correlation was 0.36 (between PRC and RF). GI and PR in particular were the least correlated to the six standard techniques, with $\tau < 0.2$.

While Fig. 1 presented the average correlations between the 17 filter-based feature selection techniques, further analysis is provided below to demonstrate how this correlation matrix varies depending on the dataset. The below-diagonal entries provide some information on variability by showing the standard deviation of each individual correlation, but this is incomplete. One option is to present all 17 correlation matrices individually, but this can be redundant and difficult to analyze. Instead, we use the Frobenius (or Hilbert–Schmidt) norm (Golub and Van Loan 1996) $\mathcal{F}$ to measure the difference between correlation matrices generated for each dataset and the "average" correlation over all 17 datasets (from Fig. 1). Let C^d be the 17 × 17 correlation matrix for dataset d. Entry $C^d_{i,j}$ is the Kendall's Tau rank correlation between filters f_i and f_j for dataset d. Let $\bar{C}$ represent the average correlation matrix over the 17 datasets, i.e., $\bar{C}_{i,j} = \frac{1}{17}\sum_{d=1}^{17} C^d_{i,j}$ in Fig. 1.

Table 3 Distance (Frobenius norm) from the mean by dataset

Dataset	Frobenius norm distance to the mean correlation matrix
ECML	3.60
CNS	2.20
Colon	1.31
Tum	1.22
Lymph	1.62
DLB	1.26
LC	2.72
ALL	1.45
Pros	1.35
MLL	1.57
Brst	2.03
AAL	0.87
Tran	4.65
Ov	3.88
NIH	4.14
Lung	2.13
Brain	2.24

The distance between correlation matrix C^d and $\bar{C}$ using the Frobenius norm $\mathcal{F}$ is calculated as:

$$\mathcal{F}_d = ||C^d - \bar{C}||_2 = \sqrt{\sum_{i=1}^{17}\sum_{j=1}^{17}\left(C^d_{i,j} - \bar{C}_{i,j}\right)^2}.$$

Since the correlation matrices are symmetric with 1 on the diagonal entries, the Frobenius norm simplifies to

$$\mathcal{F}_d = \sqrt{2\sum_{i,j=1;i<j}^{17}\left(C^d_{i,j} - \bar{C}_{i,j}\right)^2}.$$

Table 3 presents the Frobenius norm for the correlation matrix generated from each individual dataset, compared to the average correlation from Fig. 1. The dataset which generated the most dissimilar correlation matrix from the mean, as measured by the Frobenius norm, is Translation, with $\mathcal{F} = 4.65$. Figure 2 presents the correlation matrix for this dataset. Note that in Fig. 2, the below-diagonal entries are blank—since this figure considers only a single dataset, there is no standard deviation of τ. Comparing Figs. 1 and 2, there is substantially higher correlation between the six standard filters and the 11 proposed TBFS filters in Fig. 2. In particular, for this dataset, Pow, MI, KS, Dev, GM, AUC and PRC almost all have a correlation of more

	χ2	GR	IG	RF	RFW	SU	F	OR	Pow	PR	GI	MI	KS	Dev	GM	AUC	PRC
χ2		.72	.89	.31	.32	.86	.05	.43	.66	.33	.07	.84	.77	.89	.53	.76	.78
GR			.74	.20	.21	.82	.05	.48	.52	.32	.05	.68	.59	.72	.37	.59	.61
IG				.30	.31	.88	.06	.46	.66	.32	.09	.85	.76	.88	.53	.76	.78
RF					.84	.29	.02	.19	.37	.19	.11	.33	.42	.33	.54	.42	.40
RFW						.30	.05	.21	.32	.16	.10	.33	.41	.34	.51	.42	.39
SU							.06	.46	.63	.31	.08	.81	.73	.85	.50	.73	.74
F								.01	-.11	-.24	.23	.03	.04	.06	.01	.04	.01
OR									.32	.64	.18	.45	.43	.49	.31	.43	.44
Pow										.50	.07	.76	.79	.73	.65	.79	.82
PR											.20	.40	.37	.38	.34	.37	.42
GI												.10	.13	.11	.17	.13	.12
MI													.82	**.93**	.59	.82	.85
KS														.84	.73	**1.00**	**.92**
Dev															.58	.84	.86
GM																.73	.71
AUC																	**.92**
PRC																	

Fig. 2 Correlation matrix for the translation dataset

than 0.5 with χ^2, GR, IG and SU. On the other hand, the correlation relationships among the six standard filters is very similar to the average, though the pairwise relationships between χ^2, GR and IG are somewhat less strong than in the average case.

Figure 3 shows the correlation matrix among filtering techniques for the Lung dataset. From Table 3, this particular dataset has a Frobenius norm of 2.13, which is very close to the median (2.03) and average (2.25) values. The correlation structure presented in this figure is more closely aligned with that of Fig. 1. There are no correlations above 0.5 in the top right quadrant, though in general the correlations tend to be higher than average. Similar relationships are observed for the six standard filters, and generally speaking, patterns of correlation for the 11 TBFS techniques are similar to that of Fig. 1. Recall that 15 of the 17 datasets considered in this work are microarray expression datasets, with the exceptions being Translation and Ovarian. Note that in Table 3, these datasets generate two of the most divergent correlation matrices from the group average (with the highest and third-highest $\mathcal{F}$ of all datasets). One possible reason for the divergence of the two non-microarray datasets from the overall average could be because of how the filters perform on the irrelevant features. On the microarray sets, features represent genes which are active or inactive in a given cell. If a gene is not relevant to the class, it implies that this gene performs some other function in the cell which is unrelated to the class. Thus, one would expect this gene to be at equal levels in all instances, regardless of their class. There will be some variation, and the gene's level might depend on some unrelated factor which might vary significantly among examples; nonetheless, irrelevant features would be mostly constant (either low or high) across all instances. For the non-microarray sets, the features consist of either the presence of short (3 to 6 character) sequences in approximately 200 character long strings (Translation) or mass/charge values which may or may not correlate to an extant protein, let alone one pertaining to the class (Ovarian). In each of these cases, many of the irrelevant features will almost always have values of 0; if there is no underlying reason for a given sequence to be found in a string, or a given mass/charge value to describe a protein, it is very unlikely that it will be found in a given instance. Plus, even for the instances which do randomly have hits, the distribution will not be the same as for the microarray random variation. It is possible that two filters which perform equally well on the relevant features are only equivalent on the irrelevant ones when they match one of these profiles (e.g., microarray or non-microarray irrelevant features); with the other profile, new differences arise. Enough of these differences, in both directions, would explain the large values of the Frobenius norms.

We omit the presentation of the correlation matrix of the dataset with the smallest Frobenius distance from the mean correlation matrix (dataset AAL, with $\mathcal{F} = 0.87$) since it is very similar to the mean correlation matrix already shown in Fig. 1.

In Table 3, $\mathcal{F}$ varies between 0.87 and 4.65, with an average value of 2.25 and a standard deviation of 1.15. To gain a further understanding of what these distances represent, we can calculate the maximum Frobenius norm $\mathcal{F}^{\max}$ as follows. Let

	χ2	GR	IG	RF	RFW	SU	F	OR	Pow	PR	GI	MI	KS	Dev	GM	AUC	PRC
χ2		.78	**.92**	.26	.21	.67	.22	.34	.39	.24	.23	.46	.38	.47	.28	.27	.42
GR			.79	.24	.20	.64	.19	.35	.38	.25	.25	.42	.34	.43	.26	.24	.39
IG				.26	.21	.65	.23	.34	.39	.24	.23	.47	.39	.46	.29	.27	.42
RF					.67	.26	.14	.32	.39	.23	.23	.40	.32	.40	.21	.20	.40
RFW						.22	.12	.25	.30	.18	.17	.30	.24	.31	.16	.16	.31
SU							.21	.36	.40	.26	.25	.44	.36	.45	.27	.25	.41
F								.04	.10	-.06	-.06	.30	.46	.27	.55	.60	.30
OR									.56	.54	.52	.50	.28	.53	.15	.14	.50
Pow										.39	.38	.70	.43	.73	.25	.25	.69
PR											**.91**	.30	.15	.34	.06	.04	.35
GI												.29	.15	.33	.05	.03	.34
MI													.67	**.93**	.47	.45	.77
KS														.63	.73	.64	.64
Dev															.43	.42	.77
GM																.70	.47
AUC																	.50
PRC																	

Fig. 3 Correlation matrix for the Lung dataset

$$C_{i,j}^{\max} = \begin{cases} 1 & \text{if } i = j \\ 1 & \text{if } \bar{C}_{i,j} < 0.5 \quad \text{and} \quad i \neq j \\ 0 & \text{if } \bar{C}_{i,j} \geq 0.5 \quad \text{and} \quad i \neq j \end{cases}$$

In other words, for entries on the diagonal, $C_{i,j}^{\max} = 1$. For off-diagonal elements $C_{i,j}^{\max} = \arg\max(|\, x - \bar{C}_{i,j}\,|), \quad x \in \{0, 1\}$. Then

$$\begin{aligned} \mathcal{F}^{\max} &= ||C^{\max} - \bar{C}||_2 \\ &= \sqrt{\sum_{i=1}^{17}\sum_{j=1}^{17}\left(C_{i,j}^{\max} - \bar{C}_{i,j}\right)^2} \\ &= 14.22. \end{aligned}$$

In other words, the maximum distance from the mean correlation matrix $\bar{C}$ is 14.22.[1]

We would also like to understand if the Frobenius norms generated by our 17 datasets represent significant relationships, or are the correlations simply random. Forty random correlation matrices were generated, where the entries were between 0 and 1 with uniform distribution (entries on the diagonal were set to 1, and the matrix was symmetric, i.e., if entry i, j was randomly set to 0.63, then entry j, i was also set to 0.63). Then we calculated the Frobenius norm distance for each of these 40 random matrices to the mean matrix generated by our 17 datasets $\bar{C}$. Table 4 shows the Frobenius norms (vis-à-vis $\bar{C}$) for the 40 random matrices, along with the average and standard deviation. Comparing this data to that of Table 3, all of the randomly generated matrices are at substantially further distances than all 17 of the correlation matrices, where the largest Frobenius norm is 4.65. The distances for the random matrices are generally 2 or 3 times greater than those in Table 3.

We can also perform statistical tests to validate the hypothesis that the mean Frobenius norm for the random matrices $\mathcal{F}^{\text{ran}} = 0$, in other words, that in actuality the random matrices have a distance of 0 from $\bar{C}$ and that the data in Table 3 is simply due to random chance. Clearly this hypothesis is rejected since the t-statistic is $\frac{6.97}{0.36/\sqrt{40}} = 123.00$. This data suggests that the correlations among the 17 filter-based feature selection techniques is significantly different than random—in other words, the variations in correlations are not simply random but represent true relationships among the techniques.

7.2 Classification results

This section compares the 17 feature selection techniques using two different learners, Naive Bayes (NB) and Support Vector Machines (SVM). For SVM, the complexity constant `c` was changed from 1.0 to 5.0 and the `buildLogisticModels` parameter was enabled. These parameter changes were made after preliminary experimentation showed that they generally improved the performance of the SVM learner. The default parameters in Weka were used for the NB learner. Both learners were built in Weka using 10-fold cross validation. The dataset is

[1] More precisely, 14.22 is the largest distance for a correlation matrix with non-negative entries. While some correlation matrices do have negative entries, this is a relatively uncommon occurrence and we hence exclude this possibility.

Table 4 Frobenius norm distances to $\bar{C}$ for 40 randomly generated correlation matrices

6.23	6.93	7.43	6.43	7.08	6.64	7.18	6.80	7.25	6.88
6.39	6.96	7.44	6.54	7.09	6.65	7.18	6.82	7.30	6.92
6.40	6.97	7.56	6.59	7.12	6.77	7.18	6.84	7.37	7.38
6.42	7.07	7.60	6.59	7.17	6.78	7.24	6.88	7.37	7.41

Mean = 6.97, standard deviation = 0.36

partitioned into 10 disjoint, equal-sized subsets, and nine of the partitions are combined to form the training data, while the hold-out partition is used as test data. Each feature selection method is applied to the training data, and a classifier is built using the entire training dataset with a reduced set of features. The process is repeated with each partition as the test data. Further, to offset any anomalies that might be due to the cross-validation process, the entire procedure was repeated four times. The classification results were aggregated across each run of cross validation, and averaged over the four runs.

For these experiments, the "best" 25 features according to each technique were selected for classifier construction. Other values were attempted (e.g., 10, 50, 100, 1 %) but 25 generally performed well for many different datasets. Generally speaking, as the number of features increased, the relative performances of the filters converged as most of the filters select most of the important features given a large enough value for this parameter. Therefore, it is generally more useful to compare techniques with a smaller number of selected features, and since both NB and SVM performed well with 25 selected features, we present those results here.

The learners are evaluated using the area under the ROC curve, denoted AROC. This should not be confused with the TBFS filter which utilizes the area under the ROC curve, which we denote as AUC. The principals are similar, but the AROC uses the posterior probability calculated by the learner (NB or SVM) to compute the true positive and false positive rates across decision thresholds. The AROC is a commonly used measure of learner performance (Witten 2005).

Tables 5 and 6 present the results using the NB learner, while Tables 7 and 8 are for the SVM learner. First, the average AROC for each dataset is provided, followed by a summary of the performance for each filter, categorizing the datasets as "very easy", "easy", "moderate" and "difficult" to learn. Note that this categorization changes slightly depending on the learner used. Averaged over all datasets, the TBFS filter with GM obtained the highest

Table 5 Classification results by dataset, NB learner

Filter	LC	Ov	DLB	ALL	AAL	MLL	Lung	Tum	Lymph	Brain	Colon	ECML	Tran	Pros	Brst	CNS	NIH
χ^2	1	0.993	0.987	0.986	0.979	0.969	0.963	0.927	0.911	0.914	0.876	0.745	0.906	0.710	0.651	0.624	0.574
GR	0.996	0.995	0.989	0.978	0.975	0.922	0.821	0.915	0.912	0.817	0.846	0.876	0.722	0.694	0.638	0.600	0.578
IG	1	0.993	0.987	0.986	0.979	0.970	0.962	0.920	0.896	0.924	**0.880**	0.873	0.909	0.685	0.641	0.651	0.580
RF	0.987	0.988	0.991	0.985	0.968	0.963	0.960	0.968	0.851	0.766	0.844	0.873	0.764	0.664	0.733	0.573	0.573
RFW	0.987	0.989	0.988	0.986	0.969	**0.973**	0.879	**0.982**	0.864	0.546	0.857	0.809	0.756	0.713	**0.762**	0.607	0.580
SU	0.999	0.993	0.986	0.984	0.979	0.954	0.964	0.925	0.914	0.906	0.879	0.870	0.899	0.703	0.653	0.661	0.574
F	0.998	0.989	0.989	0.986	0.979	0.950	0.980	0.938	0.893	0.923	0.860	0.824	0.917	0.721	0.667	0.647	0.516
OR	0.998	0.994	0.989	0.985	0.981	0.950	0.940	0.936	0.912	0.918	0.871	0.773	0.747	0.686	0.661	0.571	0.555
Pow	0.990	0.992	0.998	0.981	0.956	0.906	0.912	0.926	0.905	0.835	0.843	0.863	0.827	0.679	0.650	0.562	0.519
PR	0.990	0.992	0.998	0.972	0.951	0.901	0.890	0.930	0.904	0.823	0.839	0.871	0.756	0.685	0.649	0.556	0.481
GI	0.992	0.993	0.984	0.973	0.975	0.906	0.891	0.928	0.900	0.836	0.839	0.853	0.758	0.679	0.649	0.554	0.496
MI	0.998	0.989	0.987	0.986	0.979	0.944	0.972	0.937	0.896	**0.934**	0.873	0.817	0.904	0.709	0.673	0.621	0.577
KS	1	0.988	0.987	0.987	0.979	0.957	**0.983**	0.934	0.888	0.923	0.863	0.942	0.916	0.746	0.686	0.634	0.564
Dev	0.999	0.989	0.983	0.986	0.979	0.940	0.970	0.934	0.875	0.908	0.879	0.860	0.899	0.729	0.674	0.603	0.537
GM	1	0.988	0.988	0.987	0.976	0.960	0.983	0.932	0.886	0.930	0.860	0.940	0.919	**0.748**	0.680	**0.664**	0.568
AUC	0.999	0.989	0.988	0.988	0.984	0.951	0.979	0.935	0.906	0.924	0.852	**0.964**	**0.921**	0.736	0.637	0.623	0.578
PRC	1	0.993	0.988	0.986	0.973	0.922	0.972	0.942	**0.924**	0.927	0.857	0.819	0.914	0.699	0.670	0.579	**0.592**
Avg	0.996	0.991	0.989	0.984	0.974	0.943	0.942	0.936	0.896	0.868	0.860	0.857	0.849	0.705	0.669	0.608	0.555
SD	0.005	0.002	0.004	0.005	0.009	0.024	0.047	0.016	0.019	0.098	0.014	0.058	0.078	0.025	0.033	0.037	0.033

Largest or smallest values for a column (or row) are shown in bold

Table 6 Summarized classification results, NB learner

Filter	Very easy	Easy	Moderate	Difficult	Avg	SD	Range	Min.
χ^2	0.989	0.953	0.870	0.640	0.866	0.145	0.426	0.574
GR	0.987	0.886	0.835	0.628	0.840	0.144	0.418	0.578
IG	0.989	0.951	0.896	0.639	0.873	0.140	0.420	0.580
RF	0.984	**0.964**	0.820	0.636	0.850	0.147	0.419	0.573
RFW	0.984	0.944	0.766	**0.666**	0.838	0.155	0.443	0.546
SU	0.988	0.948	0.894	0.648	0.873	0.137	0.426	0.574
F	0.988	0.956	0.883	0.638	0.869	0.146	0.482	0.516
OR	0.990	0.942	0.844	0.618	0.851	0.153	0.443	0.555
Pow	0.983	0.915	0.855	0.603	0.844	0.152	0.479	0.519
PR	0.981	0.907	0.839	0.593	0.835	0.157	0.517	0.481
GI	0.984	0.908	0.837	0.594	0.836	0.156	0.498	0.496
MI	0.988	0.951	0.885	0.645	0.870	0.139	0.422	0.577
KS	0.988	0.958	0.906	0.657	0.881	0.137	0.436	0.564
Dev	0.987	0.948	0.884	0.636	0.867	0.144	0.462	0.537
GM	0.988	0.958	0.907	0.665	**0.883**	**0.134**	0.431	0.568
AUC	**0.990**	0.955	**0.913**	0.644	0.880	0.143	0.422	0.578
PRC	0.988	0.946	0.888	0.635	0.868	0.144	0.421	0.579

Largest or smallest values for a column (or row) are shown in bold

Table 7 Classification results by dataset, SVM learner

Filter	LC	Ov	DLB	ALL	AAL	Lung	MLL	Tum	Pros	Tran	Brain	ECML	Colon	Lymph	Brst	CNS	NIH
χ^2	0.993	0.993	0.989	0.985	0.974	0.978	0.968	0.900	**0.936**	0.945	0.900	0.769	0.856	0.862	0.712	0.612	0.562
GR	0.998	0.999	0.990	0.976	0.987	0.866	0.933	0.905	0.907	0.672	0.866	0.872	0.850	0.836	0.641	0.576	0.581
IG	0.996	0.992	0.993	0.988	0.975	0.978	0.969	0.896	0.934	0.946	**0.917**	0.868	0.860	0.873	0.671	0.612	0.582
RF	0.995	0.999	0.979	0.971	0.972	0.967	0.970	0.969	0.917	0.790	0.776	0.921	0.834	0.854	0.744	0.484	0.578
RFW	1	1	0.984	0.969	0.971	0.941	0.970	**0.973**	0.895	0.778	0.467	0.732	0.807	0.854	**0.751**	0.600	0.596
SU	0.997	0.994	0.990	0.982	0.974	0.972	0.971	0.923	0.932	0.931	0.896	0.817	0.852	0.874	0.671	0.612	0.583
F	0.996	0.993	0.996	0.987	0.974	0.974	0.980	0.925	0.916	0.954	0.901	0.829	0.854	0.827	0.680	0.589	0.539
OR	0.998	0.997	0.990	0.982	0.978	0.960	0.974	0.914	0.915	0.728	0.914	0.807	0.863	**0.884**	0.654	0.559	0.574
Pow	0.996	1	0.981	0.969	0.984	0.941	0.906	0.899	0.909	0.877	0.880	0.878	0.848	0.851	0.661	0.578	0.513
PR	0.998	1	0.981	0.960	0.976	0.943	0.909	0.882	0.878	0.793	0.879	0.850	0.857	0.850	0.651	0.556	0.562
GI	0.998	0.997	0.985	0.973	0.979	0.943	0.896	0.882	0.873	0.797	0.905	0.850	0.851	0.850	0.638	0.556	0.495
MI	1	0.994	0.993	0.987	0.975	0.974	0.980	0.916	0.909	0.943	0.903	0.805	0.858	0.833	0.697	**0.621**	0.575
KS	0.997	0.994	0.993	0.989	0.972	0.977	0.972	0.915	0.911	0.953	0.905	**0.964**	0.855	0.808	0.714	0.604	0.577
Dev	0.997	0.994	0.990	0.988	0.976	0.975	0.976	0.918	0.905	0.942	0.891	0.858	0.861	0.833	0.691	0.604	0.595
GM	0.998	0.995	0.996	0.993	0.975	0.979	0.984	0.912	0.907	**0.956**	0.903	0.958	0.848	0.820	0.713	0.596	0.564
AUC	0.997	0.995	0.996	0.988	0.976	0.978	0.975	0.925	0.902	0.954	0.905	0.941	**0.863**	0.863	0.722	0.604	0.589
PRC	0.995	0.995	0.989	0.986	0.972	0.969	0.964	0.918	0.916	0.950	0.892	0.796	0.853	0.868	0.751	0.621	**0.598**
Avg	0.997	0.996	0.989	0.981	0.976	0.960	0.959	0.916	0.910	0.877	0.865	0.854	0.851	0.849	0.692	0.587	0.569
SD	0.002	0.003	0.005	0.009	0.004	0.028	0.029	0.025	0.017	0.095	0.107	0.065	0.013	0.021	0.038	0.034	0.029

Largest or smallest values for a column (or row) are shown in bold

AROC (0.883) with the lowest variance for the NB learner. For the SVM learner, the TBFS filter with AUC obtained the highest AROC (0.893) with the second lowest variance.

There is relatively little differentiation among filters for the very easy datasets. In general, as the datasets become harder to learn, the variation between techniques increases. The optimal filter does vary greatly across datasets, but the preferred filter should perform well in a wide variety of situations and rarely perform very poorly. Therefore, we seek a filter that obtains a high AROC with low variance. For both the NB and SVM learners, the TBFS filters using AUC, KS, GM and PRC generally exhibit these qualities. Of the standard filters, for the SVM learner, IG obtains the

Table 8 Summarized classification results, SVM learner

Filter	Very easy	Easy	Moderate	Difficult	Avg	SD	Range	Min.
χ^2	0.983	0.918	0.866	0.629	0.878	0.136	0.431	0.562
GR	0.964	0.906	0.819	0.599	0.850	0.145	0.423	0.576
IG	0.985	0.915	0.893	0.622	0.885	0.135	0.414	0.582
RF	0.979	**0.943**	0.835	0.602	0.866	0.151	0.515	0.484
RFW	0.976	0.934	0.728	0.649	0.840	0.165	0.533	0.467
SU	0.983	0.928	0.874	0.622	0.881	0.135	0.415	0.583
F	0.986	0.921	0.873	0.603	0.877	0.145	0.457	0.539
OR	0.983	0.915	0.839	0.596	0.864	0.148	0.439	0.559
Pow	0.968	0.904	0.867	0.584	0.863	0.144	0.487	0.513
PR	0.967	0.880	0.846	0.590	0.854	0.140	0.444	0.556
GI	0.967	0.877	0.851	0.563	0.851	0.152	0.503	0.495
MI	0.986	0.912	0.868	0.631	0.880	0.134	0.424	0.575
KS	0.985	0.913	0.897	0.632	0.888	0.135	0.421	0.577
Dev	0.985	0.912	0.877	0.630	0.882	0.132	0.401	0.595
GM	**0.989**	0.910	0.897	0.624	0.888	0.139	0.433	0.564
AUC	0.986	0.913	**0.905**	0.638	**0.893**	0.131	0.408	0.589
PRC	0.982	0.917	0.872	**0.657**	0.884	**0.125**	0.397	0.598

Largest or smallest values for a column (or row) are shown in bold

highest average performance with the lowest volatility, while SU is preferred for the NB learner. RFW was often the best overall technique for some datasets, but for others, it performed very poorly. From these results, we conclude that the TBFS filter often performs better than the standard filters with lower volatility, and in particular we recommend the TBFS filter used with AUC or GM.

8 Conclusions

This work proposes 11 new threshold-based feature selection techniques for ranking attributes based on the strength of the relationship between the attribute and the class. Each attribute is paired individually with the class and normalized so that its values range from 0 to 1. A decision threshold t is then used to categorize examples as either positive or negative based on the value of the normalized attribute. As the decision threshold is varied from 0 to 1, the categorization of examples changes. This is analogous to the decision threshold adjustment for a posterior probability calculated by a classifier. As t changes, a variety of metrics can be calculated, for example the false positive, true positive, false negative and true negative rates. For 7 of the 11 metrics (GM, F, OR, Pow, PR, MI and KS), the maximum value over all possible decision thresholds is utilized. For two metrics (GI and Dev) the minimum value is used. For the remaining two metrics (AUC and PRC), the area under a curve is calculated. Therefore, each feature in a dataset will have a "score", representing the strength of the association with the class, relative to each metric. Once these scores are obtained, the attributes can be ranked from most predictive to least predictive using a single TBFS filter. For example, a practitioner may choose to select the 10 % most predictive attributes as determined by the AUC TBFS filter. This reduced dataset can then be used for any type of domain-specific analysis, e.g., classifier construction, clustering.

When proposing new feature ranking techniques, it is important to understand how different the techniques are from one another and from currently used techniques. If the proposed techniques generate feature rankings that are similar to the feature rankings of currently used feature selection techniques, then these new techniques provide little insight. To achieve this objective, Kendall's Tau rank correlation, a non-parametric measure of correlation between two feature rankings, is used. Many studies only construct classifiers after using the various feature selection techniques to determine which technique is "best". Using this approach alone has shortcomings, however—for example, it is unclear which classifier(s) should be used. If two feature selection techniques produce similar results using a single classifier, it is unclear if it is because both feature selection techniques identify the exact same set of features or because the classifier is able to generalize well from two different sets of features and in fact the underlying sets of selected features are quite different. Will this pattern hold if other classifiers are used? Further, the ultimate objective of feature selection may be something besides classification. For example, practitioners are often interested in knowing which features are most predictive (and conversely, which are least predictive). Our experimental analysis, therefore, combines both evaluation approaches.

In addition to comparing the feature rankings directly, the classification results for two learners, Naive Bayes and Support Vector Machines, are also presented.

The experimental results demonstrate that the newly proposed techniques differ significantly from those of the standard filter-based feature selection techniques used in our study. The correlations between the 11 TBFS techniques are generally low to moderate. In addition, some of the TBFS filters (AUC, GM, KS, PRC) performed very well when utilized with both the NB and SVM learners, often outperforming the six standard filters. Therefore, we conclude that these new procedures represent a valuable addition to the repository of filter-based feature selection techniques that are currently available to practitioners analyzing high dimensional bioinformatics datasets.

The TBFS method can be easily extended to additional metrics in future work. Experimental evaluation with additional datasets, both in the bioinformatics domain and in other high-dimensional domains such as text mining, should also be considered. Future work should also evaluate the robustness and stability of the newly proposed TBFS techniques, as done in previous work for other filters (Kalousis et al. 2007; Křížek et al. 2007).

Appendix

Kolmogorov–Smirnov statistic

The Kolmogorov–Smirnov statistic (KS) (Conove 1971) measures the maximum difference between the cumulative distribution functions of examples in each class based on the normalized attribute $\hat{X}^j$. The distribution function $F_c(t)$ for a class c is estimated by the proportion of examples x from class c with $\hat{X}^j(x) \le t, \quad 0 \le t \le 1$. In a two class setting with $c \in \{N, P\}$ we have

$$F_P(t) = \frac{|\{x \in D | (\hat{X}^j(x) \le t) \cap (c(x) = P)\}|}{|\{x \in D | c(x) = P\}|},$$

$$F_N(t) = \frac{|\{x \in D | (\hat{X}^j(x) \le t) \cap (c(x) = N)\}|}{|\{x \in D | c(x) = N\}|}.$$

KS is computed as

$$\mathrm{KS} = \max_{t \in [0,1]} |F_P(t) - F_N(t)|.$$

With respect to KS, an attribute provides the best performance at a specific t value when the distance between the two distribution functions is maximized. The larger the KS value, the better the attribute is able to separate the two classes, and hence the more significant the attribute is. An attribute that perfectly separates examples of the two classes obtains a KS of 1 while an attribute that is unable to separate the two classes provides a KS of 0.

Alternatively, the KS can also be calculated as the maximum difference between the curves generated by the true positive and false positive rates (TPR(t) and FPR(t)) as the decision threshold changes from 0 to 1. $F_P(t)$ can be further defined as follows:

$$\begin{aligned} F_P(t) &= \frac{|\{x \in D | (\hat{X}^j(x) \le t) \cap (c(x) = P)\}|}{|\{x \in D | c(x) = P\}|} \\ &= \frac{|\{x \in D | (\hat{c}^t(x) = N) \cap (c(x) = P)\}|}{|\{x \in D | c(x) = P\}|} \\ &= \frac{\mathrm{FN}(t)}{|\{x \in D | c(x) = P\}|} = \mathrm{FN}R(t) = 1 - \mathrm{TPR}(t). \end{aligned}$$

Similarly, it can be shown that $F_N(t) = 1 - \mathrm{FPR}(t)$, and it therefore follows that

$$\mathrm{KS} = \max_{t \in [0,1]} |\mathrm{TPR}(t) - \mathrm{FPR}(t)|.$$

Gini index

The Gini index (GI) was first introduced by Brieman et al. (1984) within the CART algorithm. For a given threshold t, let $S_t = \{x \mid \hat{X}^j(x) > t\}$ and $\bar{S}_t = \{x \mid \hat{X}^j(x) \le t\}$. Then the Gini index is calculated as:

$$\begin{aligned} \mathrm{GI} &= \min_{t \in [0,1]} \left[1 - \left(P^2(\mathrm{TP}(t) \mid S_t) + P^2(\mathrm{FP}(t) \mid S_t)\right)\right] \\ &\quad + \left[1 - \left(P^2(\mathrm{TN}(t) \mid \bar{S}_t) + P^2(\mathrm{FN}(t) \mid \bar{S}_t)\right)\right] \\ &= \min_{t \in [0,1]} \left[1 - \left(\left(\frac{\mathrm{TP}(t)}{\mathrm{TP}(t) + \mathrm{FP}(t)}\right)^2 + \left(\frac{\mathrm{FP}(t)}{\mathrm{TP}(t) + \mathrm{FP}(t)}\right)^2\right)\right] \\ &\quad + \left[1 - \left(\left(\frac{\mathrm{TN}(t)}{\mathrm{TN}(t) + \mathrm{FN}(t)}\right)^2 + \left(\frac{\mathrm{FN}(t)}{\mathrm{TN}(t) + \mathrm{FN}(t)}\right)^2\right)\right] \\ &= \min_{t \in [0,1]} \left[1 - \left(\mathrm{PRE}(t)^2 + (1 - \mathrm{PRE}(t))^2\right)\right] \\ &\quad + \left[1 - \left(\mathrm{NPV}(t)^2 + (1 - \mathrm{NPV}(t))^2\right)\right] \\ &= \min_{t \in [0,1]} [2\mathrm{PRE}(t)(1 - \mathrm{PRE}(t)) + 2\mathrm{NPV}(t)(1 - \mathrm{NPV}(t))]. \end{aligned}$$

NPV or *negative predicted value* represents the percentage of examples predicted to be negative that are actually negative (Weiss 2003) and is very similar to the precision—in fact, it is often thought of as the precision of instances predicted to be in the negative class. The Gini index for the attribute is then the minimum Gini index at all decision thresholds $t \in [0, 1]$.

References

Akbani R, Kwek S (2005) Adapting support vector machines to predict translation initiation sites in the human genome. pp 143–145

Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, Levine AJ (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc Nal Acad Sci USA 96(12): 6745–6750

Au W-H, Chan KCC, Wong AKC, Wang Y (2005) Attribute clustering for grouping, selection, and classification of gene expression data. IEEE/ACM Trans Comput Biol Bioinform 2(2):83–101

Battiti R (1994) Using mutual information for selecting features in supervised neural net learning. IEEE Trans Neural Netw 5(4): 537–550

Breiman L, Friedman J, Olshen R, Stone C (1984) Classification and regression trees. Chapman and Hall/CRC Press, Boca Raton, FL

Brown MPS, Grundy WN, Lin D, Cristianini N, Sugnet CW, Furey TS, Ares M, Haussler (2000) Knowledge-based analysis of microarray gene expression data by using support vector machines. Proc Natl Acad Sci USA 97(1):262–267

Chen X-w, Wasikowski M (2008) Fast: a ROC-based feature selection metric for small samples and imbalanced data classification problems. In: Proceedings of 14th ACM SIGKDD international conference on knowledge discovery and data mining KDD '08, pp 124–132, ACM, New York, NY, USA

Conover WJ (1971) Practical nonparametric statistics, 2nd edn. Wiley, New York

Datta S, Datta S (2003) Comparisons and validation of statistical clustering techniques for microarray gene expression data. Bioinformatics 19(4):459–466

Davis CA, Gerick F, Hintermair V, Friedel CC, Fundel K, Küffner R, Zimmer R (2006) Reliable gene signatures for microarray classification: assessment of stability and performance. Bioinformatics 22(19):2356–2363

Ding C, Peng H (2003) Minimum redundancy feature selection from microarray gene expression data. In: Proceedings of the IEEE computer society conference on bioinformatics, CSB '03, IEEE Computer Society, Washington, DC, USA, p 523

Fayyad UM, Irani KB (1992) On the handling of continuous-valued attributes in decision tree generation. Mach Learn 8: 87–102

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

Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, Haussler D (2000) Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics 16(10):906–914

Golub GH, CF Van Loan (1996) Matrix computations, 3rd edn. The Johns Hopkins University Press, Baltimore

Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286(5439):531–537

Gupta R, Mittal A, Singh K (2008) A novel and efficient technique for identification and classification of gpcrs. IEEE Trans Inf Technol Biomed 12(4):541–548

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

Hall MA, Holmes G (2003) Benchmarking attribute selection techniques for discrete class data mining. IEEE Trans Knowl Data Eng 15(6):392–398

Hall MA, Smith LA (1999) Feature selection for machine learning: Comparing a correlation-based filter approach to the wrapper. In: Proceedings of the twelfth international Florida Artificial Intelligence Research Society conference, pp 235–239

Petricoin EF III, Ardekani AM, Hitt BA, Levine PJ, Fusaro VA, Steinberg SM, Mills GB, Simone C, Fishman DA, Kohn EC, Liotta LA (2002) Use of proteomic patterns in serum to identify ovarian cancer. Lancet 359(9306):572–577

Inza I, Larraaga P, Blanco R, Cerrolaza AJ (2004) Filter versus wrapper gene selection approaches in dna microarray domains. Artif Intell Med 31(2):91–103 (Data Mining in Genomics and Proteomics)

Jirapech-Umpai T, Aitken S (2005) Feature selection and classification for microarray data analysis: evolutionary methods for identifying predictive genes. BMC Bioinform 6(1):148

Kalousis A, Prados J, Hilario M (2007) Stability of feature selection algorithms: a study on high-dimensional spaces. Knowl Inf Syst 12(1):95–116

Kamal AHM, Zhu X, Pandya A, Hsu S, Shoaib M (2009) The impact of gene selection on imbalanced microarray expression data. In: Proceedings of 1st international conference on bioinformatics and computational biology, BiCoB '09, April 2009, Berlin/ Heidelberg, Springer, pp 259–269

Khoshgoftaar TM, Allen EB, Deng J (2002) Using regression trees to classify fault-prone software modules. IEEE Trans Reliab 51(4): 455–462

Kira K, Rendell LA (1992) The feature selection problem: Traditional methods and a new solution. In: Proceedings of 10th national conference on artificial intelligence, AAAI '92, no 10, pp 129–134. Wiley, New York

Kononenko I (1994) Estimating attributes: analysis and extensions of RELIEF. In: European conference on machine learning, Springer, New York, pp 171–182.

Kuncheva LI (2007) A stability index for feature selection. In: Proceedings of the 25th conference on proceedings of the 25th IASTED international multi-conference, AIAP'07, ACTA Press, Anaheim, CA, USA, pp 390–395

Křížek P, Kittler J, Hlaváč V (2007) Improving stability of feature selection methods. In: Proceedings of the 12th international conference on computer analysis of images and patterns, CAIP'07, Berlin, Heidelberg, Springer, pp 929–936

Lee BJ, Lee HG, Lee JY, Ryu KH (2007) Classification of enzyme function from protein sequence based on feature representation, pp 741–747

Li L, Tang H, Wu Z, Gong J, Gruidl M, Zou J, Tockman M, Clark RA (2004) Data mining techniques for cancer detection using serum proteomic profiling. Artif Intell Med 32(2):71–83

Li L, Weinberg CR, Darden TA, Pedersen LG (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(12):1131–1142

Model F (2001) Feature selection for dna methylation based cancer classification. Bioinformatics 17:157–164(8)

Peddada SD, Lobenhofer EK, Li L, Afshari CA, Weinberg CR, Umbach DM (2003) Gene selection and clustering for time-course and dose-response microarray experiments using order-restricted inference. Bioinformatics 19(7):834–841

Pedersen AG, Nielsen H (1997) Neural network prediction of translation initiation sites in eukaryotes: perspectives for EST and genome analysis. In: Proceedings of the 5th international conference on intelligent systems for molecular biology, AAAI Press, pp 226–233

Peng H, Long F, Ding C (2005) Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell 27(8):1226–1238

Piatetsky-Shapiro G, Tamayo P (2003) Microarray data mining: facing the challenges. SIGKDD Explor Newsl 5(2):1–5

Quinlan JR (1993) C4.5: Programs for machine learning. Morgan Kaufmann, San Mateo, CA

Radivojac P, Chawla NV, Dunker AK, Obradovic Z (2004) Classification and knowledge discovery in protein databases. J Biomed Inform 37(4):224–239 Biomedical Machine Learning.

Saeys Y, Abeel T, Van de Peer Y (2008) Towards robust feature selection techniques. In: Proceedings of Benelearn, pp 45–46

Seliya N, Khoshgoftaar TM, Van Hulse J (2009) A study on the relationships of classifier performance metrics. In: Proceedings of the 21st IEEE international conference on tools with artificial intelligence (ICTAI 2009), pp 59–66, Newark, NJ

Sun Y, Robinson M, Adams R, te Boekhorst R, Rust A, Davey N (2006) Using feature selection filtering methods for binding site predictions, vol 1, pp 566–571

Van Hulse J, Khoshgoftaar TM, Napolitano A (2011) A comparative evaluation of feature ranking methods for high dimensional bioinformatics data. In: Proceedings of the 12th IEEE international conference on information reuse and integration (IRI 2011), pp 315–320, Las Vegas, NV

Van Hulse J, Khoshgoftaar TM, Napolitano A, Wald R (2009) Feature selection with high-dimensional imbalanced data. In: Proceedings of the 9th IEEE international conference on data mining—workshops (ICDM'09), Miami, FL, December 2009, IEEE Computer Society, pp 507–514

Wang Y, Makedon FS, Ford JC, Pearlman J (2005) HykGene: a hybrid approach for selecting marker genes for phenotype classification using microarray gene expression data. Bioinformatics 21(8):1530–1537

Wang Y, Tetko IV, Hall MA, Frank E, Facius A, Mayer KF, Mewes HW (2005) Gene selection from microarray data for cancer classification—a machine learning approach. Comput Biol Chem 29(1):37–46

Weiss GM, Provost F (2003) Learning when training data are costly: the effect of class distribution on tree induction. J Artif Intell Res (19):315–354

Witten IH, Frank E (2005) Data mining: practical machine learning tools and techniques, 2nd edn. Morgan Kaufmann

Xing EP, Jordan MI, Karp RM (2001) Feature selection for high-dimensional genomic microarray data. In: Proceedings of the eighteenth international conference on machine learning, ICML '01, Morgan Kaufmann, San Francisco, CA, USA, pp 601–608

Yu L, Liu H (2004) Redundancy based feature selection for microarray data. In: Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining, KDD '04, ACM, New York, NY, USA, pp 737–742

Zhang H, Yu C-Y, Singer B, Xiong M (2001) Recursive partitioning for tumor classification with gene expression microarray data. Proc Natl Acad Sci USA 98(12):6730–6735