BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 8 2012, pages 1151–1157
doi:10.1093/bioinformatics/bts108

Data and text mining

Advance Access publication March 5, 2012

Combining multiple approaches for gene microarray classification

Loris Nanni[1,*], Sheryl Brahnam[2] and Alessandra Lumini[3]

[1]Department of Information Engineering, University of Padua, Via Gradenigo, 6 – 35131, Padova, Italy, [2]Department of Computer Information Systems, Missouri State University, 901 S. National, Springfield, MO 65804, USA and [3]DEIS, Università di Bologna, Via Venezia 52, 47521 Cesena, Italy

Associate Editor: Martin Bishop

ABSTRACT

Motivation: The microarray report measures the expressions of tens of thousands of genes, producing a feature vector that is high in dimensionality and that contains much irrelevant information. This dimensionality degrades classification performance. Moreover, datasets typically contain few samples for training, leading to the 'curse of dimensionality' problem. It is essential, therefore, to find good methods for reducing the size of the feature set.

Results: In this article, we propose a method for gene microarray classification that combines different feature reduction approaches for improving classification performance. Using a support vector machine (SVM) as our classifier, we examine an SVM trained using a set of selected genes; an SVM trained using the feature set obtained by Neighborhood Preserving Embedding feature transform; a set of SVMs trained using a set of orthogonal wavelet coefficients of different wavelet mothers; a set of SVMs trained using texture descriptors extracted from the microarray, considering it as an image; and an ensemble that combines the best feature extraction methods listed above. The positive results reported offer confirmation that combining different features extraction methods greatly enhances system performance. The experiments were performed using several different datasets, and our results [expressed as both accuracy and area under the receiver operating characteristic (ROC) curve] show the goodness of the proposed approach with respect to the state of the art.

Availability: The MATHLAB code of the proposed approach is publicly available at bias.csr.unibo.it/nanni/micro.rar

Contact: loris.nanni@unipd.it

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on 22 November 2011; revised on 10 February 2012; accepted on 24 February 2012

1 INTRODUCTION

DNA microarray technology has proven to be an important breakthrough in molecular biology. This rapidly maturing technology is providing scientists with a means of monitoring the expression of genes on a genomic scale (Chee *et al*., 1996). One important application area is disease prognostication (Golub *et al*., 1999; Peng, 2006). Benefits include the potential for identifying individual genes responsible for disease (Der *et al*., 1998; Huang and Keoman, 2005; Maglietta *et al*., 2007; Turashvilli *et al*., 2007) and for providing scientists with a more accurate means of diagnosis and prognosis (Alon *et al*., 1999; Beer *et al*., 2002; Ben-Dor *et al*., 2003; Brown *et al*., 2000; Freije *et al*., 2004; Petricoin *et al*., 2002; Pomeroy *et al*., 2002; Singh *et al*., 2002; Tamayo *et al*., 1999). Large-scale profiling of gene expression can reveal, for example, normal versus malignant cells and the genetic and cellular changes in the progression of tumor metastasis (Golub *et al*., 1999).

The benefits offered by simultaneously monitoring tens of thousands of genes, however, depend on developing tools capable of handling not only the sheer size of this data but also the small number of samples usually available for analysis. Machine learning systems are well suited for this problem, but they must be designed to handle high levels of noise, as only a small minority of genes is typically relevant for any given problem. The small sample size compared to the large number of features means that these systems must also contend with the dreaded 'curse of dimensionality' (Lee *et al*., 2008). It would be very beneficial, therefore, if good methods for identifying these small sets of relevant genes could be developed.

In the literature, gene selection methods have been organized into three categories: filter, wrapper and embedded methods (Bontempi, 2007). Filter methods reveal dependencies without using classifiers and are based on statistical methods of ranking genes, e.g. *t*-statistics (Devore and Peck, 1997; Tibshirani *et al*., 2002), class separability (Dudoit *et al*., 2002) and Fisher's criterion (Broet *et al*., 2004; Lai *et al*., 2004). Wrapper and embedded methods consider the mutual information among genes as well as its relevance (Peng *et al*., 2005). Example classifiers used in wrapper methods include Bayesian classifier (Figuiredo and Jain, 2001; Hastie *et al*., 2009), K-nearest neighbor (Hastie *et al*., 2009; Tibshirani *et al*., 2003) and support vector machines (SVMs) (Furey *et al*., 2000; Guyon *et al*., 2002). Wrapper methods are much slower than filter methods because they search for optimal combinations of features/genes, but filter methods may not select the most optimal set of features.

Examples of embedded methods include one-norm SVM (Fung and Mangasarian, 2000), logistic regression (Shen and Tan, 2005), sparse logistic regression (Roth, 2004) and methods based on regularization (Ghosh and Chinnaiyan, 2005). An interesting embedded method is that developed by Huerta *et al.* (2010). They devised a Genetic Algorithm with Fisher's Linear Discriminant Analysis (LDA) as the fitness function that performed well across a number of databases using a small number of selected genes. Most of these filter, wrapper and embedded methods are comparable in accuracy (Ghorai *et al*., 2011).

Several recent advances include reducing the sample set (Chen and Lin, 2011), using classifier ensembles (Ghorai *et al*., 2011; Huang *et al*., 2010; Tan and Gilbert, 2003), rather than single classifiers and using hybrid or multiple sets of different type of

*To whom correspondence should be addressed.

feature selection and transformation methods (Ghorai *et al*., 2011). Chen and Lin (2011) have improved classifier performance by extracting significant samples that are located only on support vectors.

Huang *et al.* (2010) improve performance using decision forest for classification of gene expression data, and Stiglic *et al.* (2010) use rotation forests for robust and improved classification accuracy. Ghorai *et al.* (2011) have developed an ensemble that combines both filter and wrapper methods: a ranking method performs a fast reduction in dimensionality and a wrapper method refines the search. Their method has demonstrated comparable performance with wrapper methods while providing significant reduction in the computational burden.

In this article, we propose to classify DNA microarray data using an ensemble of SVM classifiers, with each SVM trained on a different set of features. SVM is selected because it is considered to be one of the most powerful classifiers in microarray classification of cancers (Statnikov *et al*., 2008) and in several other bioinformatic problems (Hayat and Khan, 2011; Tahir *et al*., 2011). Even though SVM is a strong learner and thus not typically suitable for ensembles, it actually performs well in ensembles if coupled with the random subspace technique (Nanni and Lumini, 2011).

In our experiments, we specifically investigate approaches: (i) that compare standard feature selection methods, where only a subset of the whole gene set is retained and then used to train an SVM; (ii) that compare several feature transform methods, where the dimension of the feature vector is reduced and then used to train an SVM; (iii) that train a set of SVMs using a set of orthogonal wavelet coefficients of different wavelet mothers-these sets of coefficients are selected via Sequential Forward Floating Selection (SFFS) using the leave-one-dataset-out validation protocol, such that when a given dataset is classified, the sets of coefficients are selected by SFFS using the others datasets as validation set; and (iv) that consider the microarray as an image, where the texture descriptors are extracted from the image and used to train an SVM. Experiments are carried out on several datasets, and experimental results show that the proposed method performs well when considering both accuracy and the area under the ROC curve (AUC) as the performance indicators.

2 METHODS

In this section, we briefly describe the feature selection, feature transform and classification and fusion methods, including the tree wavelet and texture descriptors used in our approach.

2.1 Feature selection

The feature selection methods we explore are the following:

- *Fisher score (Fi)*: a method utilizing discriminative methods and generative statistical models for determining the most relevant features for classification;
- *Gini index (Gi)*: a statistical measure of dispersion, most commonly used to quantify wealth distributions based on the Lorentz curve;
- *mRMR (Mr)*: a feature selection method that correlates the strongest features with a classification variable: features are selected that are mutually different from each other while still maintaining high correlation;
- *T-test (Tt)*: a statistical hypothesis that uses the Student's distribution; and
- *Sb*: a feature selection method based on the sparse Bayesian multinomial logistic regression.

In most cases, the code for the above methods was taken from the MATLAB Feature Selection Package available at http://featureselection.asu.edu/

In addition to the above listed feature selection methods, we also examine:

- *FFacsa2*[1] (Luo *et al*., 2011): a forward feature selection algorithm that is based on the aggregation of classifiers generated by a single attribute;
- *SVMrfe1* (Guyon *et al*., 2002): the famous SVM-based recursive feature elimination method;
- *SFFS* (Pudil *et al*., 1994)[2]: an exhaustive search procedure that has been studied extensively and shown to perform well compared to competing methods (Kudo and Sklansky, 2000). To reduce computation time, SFFS starts from the 500 genes selected by *Fi*; then the best set is extracted. SVM is used as the objective performance method.

See the Supplementary Material for a fuller discussion of Fisher score and SFFS.

2.2 Feature transform

We explore the following feature transform techniques:

- Locally Linear Embedding (LLE), as proposed in Roweis and Saul (2000);
- Orthogonal LDA (OLDA), as proposed in Ye (2005);
- Orthogonal Neighborhood Preserving Projections (ONPPs), as proposed in Kokiopoulou and Saad (2005); and
- Neighborhood Preserving Embedding (NPE), as proposed in (He *et al*., 2005). Unlike principal component analysis (PCA), which aims at preserving the global Euclidean structure of the data, NPE preserves the local neighborhood structure on the data manifold. As a result, NPE is less sensitive to outliers than is PCA. We used the MATLAB code freely available at http://www.zjucadcg.cn/dengcai/Data/data.html

See the Supplementary Material for a fuller discussion of NPE.

2.3 Tree wavelet

In the case of one dimensional wavelet decomposition, the first step produces two sets of coefficients from the signal: (i) approximation coefficients, or *scaling coefficients*; and (ii) detail coefficients, or *wavelet coefficients*. The approximation coefficients are split into two parts repeating the same algorithm, being thereby replaced by approximation coefficients and detail coefficients. This decomposition process is repeated until a required level is reached (Liu, 2009; Nanni and Lumini, 2011).

In this article, we examine the following wavelets (until the sixth decomposition level): Haar, Daubechies order 7, Symmlet order 2, Coiflets order 2, Biorthogonal order for reconstruction 2 and for decomposition 2, Reverse Biorthogonal order for reconstruction 2 and for decomposition 2. For each set of coefficients (both approximation coefficients and detail coefficients) of a given decomposition level, a different classifier is trained. The decomposition is applied both on the original data and on the set of genes selected by Fisher score. SFFS is used to select a set of subbands. The testing protocol was the leave-one-dataset-out validation protocol. When a given dataset is classified, the sets of coefficients are selected using as the validation set the other datasets. A fuller discussion of wavelet decomposition and the set of subbands selected by SFFS considering all the datasets are reported in the Supplementary Material.

[1]The MATLAB code was shared by the original authors of *FFacsa2*, which also shared the code of *SVMrfe.*

[2]Implemented as in PRTools (prtools.org/prtools.html).

2.4 Texture descriptors

In this approach, we consider the microarray as an image from which a set of texture descriptors is extracted. First, we select a set of 900 genes using the Fisher criterion feature selection method. Then this 900-dimensional feature vector is reshaped as a matrix using random assignment. A total of 50 different random reshapings are performed. For each reshaping, a different SVM is trained, with results combined using a fusion rule.

In this article, we examine the following image texture feature transforms:

- Lu is a concatenation of the uniform bins extracted using local binary patterns (LBPs) (Ojala *et al*., 2002) with $P = 8$ and $P = 16$. If $x = 8$ then $R = 1$, if $x = 16$ then $R = 2$. The length of the feature vector is 59 in the case $x = 8$ and 243 in the case $x = 16$;
- Lr is rotation invariant uniform bins extracted using LBP with $P = 8$ and $P = 16$. If $x = 8$ then $R = 1$, if $x = 16$ then $R = 2$. The length of the feature vector is 10 in the case $x = 8$ and 18 in the case $x = 16$;
- LP(x) is local phase quantization (Ojansivu and Heikkila, 2008) with radius $x = 3$ or $x = 5$. The length of the feature vector is 256 in both cases;
- LQPu is different local quinary patterns (Nanni *et al*., 2010) with uniform bins and with $\tau 1 = \{1,3,5,7,9\}$ and $\tau 2 = \{\ \tau 1+2,\ \tau 1+4,\ldots,\ \tau 1+11\}$. These are combined by a fusion rule (see the 'Results' section for details).

See the Supplementary Material for a fuller discussion of local phase quantization.

2.5 Classification and fusion

In this approach, we use SVM as the stand-alone classifier. SVM is a general purpose binary classifier based on statistical learning. It performs classification in two steps. In the first step, it maps the sample data vector into a higher dimensional data space by means of polynomial kernels or radial basis function kernels. In the second step, the algorithm finds a hyperplane in this space that has the largest margin separating the classes.

The fusion step is performed by means of the sum rule or the majority voting (vote) rule. The first consists in summing the scores of all the classifiers of the ensemble and selecting the class with the highest score; the second simply selects the class with the higher number of votes (see the Supplementary Material for a fuller discussion of SVM and the sum and vote rules).

3 RESULTS

To assess the performance of our approach, we have conducted several experiments on a number of publicly available datasets. Below we provide a brief description of each dataset (the salient features of each dataset are summarized in Table 1):

- Breast dataset (B) (van 't Veer *et al*., 2002): the goal of this experiment is to identify patients who might benefit from adjuvant chemotherapy. Two classes are considered: patients who continued to be disease free after 5 years (44 samples) and patients who developed metastases within 5 years (34 samples);
- Ovarian dataset (O) (Petricoin *et al*., 2002): the goal of this experiment is to identify proteomic patterns in serum that distinguish between ovarian cancer and normal non-cancer groups. Two classes are considered: 91 controls (Normal) and 162 ovarian cancers;

Table 1. Characteristics of the datasets used in the experiments: the first column presents the number of attributes (#A), and the second column reports the number of examples (#E)

Dataset	#A	#E
Ovarian (O)	15 154	253
Prostate (P)	12 600	102
Lung (L)	12 533	181
Breast (B)	24 481	78
Medulloblastoma (M)	7129	60
Colon (C)	2000	62
Duke (D)	7129	44
ALML (A)	7129	72
DBCL (DL)	7129	77

- Lung dataset (L)[3] (Gordon *et al*., 2002): the goal of this experiment is to classify between malignant pleural mesothelioma (MPM) and adenocarcinoma (ADCA) of the lung. Two classes are considered: 31 MPM tissue samples and 150 ADCA tissue samples;
- Prostate tumors (P) (Singh *et al*., 2002): the goal of this experiment is to classify prostate tumor samples and normal non-tumor samples. Two classes are considered: 52 prostate tumor samples and 50 normal samples;
- Medulloblastoma (M) (Pomeroy *et al*., 2002): the researchers analyze 60 similarly treated patients from whom biopsies were obtained before receiving treatment. Using this dataset, Pomeroyet *et al*. show that the clinical outcome of children with medulloblastomas is predictable on the basis of the gene expression profiles of their tumors at diagnosis;
- Colon (C) (Alon *et al*., 1999): the colon dataset contains 62 samples: 40 are tumor samples and 22 are normal controls. In this dataset, 2000 genes with highest intensity across the samples are considered;
- Duke (D) (Luo *et al*., 2011): this is a dataset that contains 44 patterns described by 7129 genes;
- ALML (A) (Golub *et al*., 1999): this leukemia dataset was derived from a study of gene expression in two types of acute leukemia: acute lymphoblastic leukemia (ALL) and acute myeloid-leukemia (AML). The dataset includes 47 cases of ALL and 25 cases of AML, together with 7129 genes;
- DLBCL (DL) (Shipp *et al*., 2002): the goal of this dataset is to distinguish diffuse large B-cell lymphoma (DLBCL) from follicular lymphoma (FL) morphology. This dataset contains 58 DLBCL samples and 19 FL samples.

In our first experiment, we compare several feature selection methods using a stand-alone SVM as the classifier for the function of the number of g genes retained: 150, 300 and 450, respectively. In Table 2, we report the average performance of the different approaches across all datasets (the performance for each dataset is reported in Supplementary Table S1 in the Supplementary Material).

[3]Publically available at http://www.chestsurg.org.

Table 2. Average accuracy obtained using different feature selection methods as a function of the number *g* genes retained

avg(ACC)		Fi	Gi	Mr	Sb	Tt	FFacsa	SVMrfe	SFFS
g	***150***	87.50	78.47	84.44	80.61	84.39	87.26	88.77	88.85
	300	87.97	83.24	87.14	80.61	85.45	86.76	88.95	88.10
	450	**89.59**	85.53	87.25	80.61	86.24	87.67	89.48	88.48

The bold values are the highest performance, the italic values are the values of parameters.

Table 3. Average accuracy obtained using different feature transform methods in reduced spaces of different dimensionality *k*

avg(ACC)		LLE	OLDA	ONPP	NPE
k	***20***	86.25	89.06	86.60	**89.93**
	30	86.42	89.06	87.40	**89.93**
	45	86.25	89.06	87.69	**89.93**

The bold values are the highest performance, the italic values are the values of parameters.

It is interesting to note that the best performance is obtained by the old Fisher criterion, which slightly outperforms the more recent *FFacsa2*, *SVMrfe* and the computationally heavy *SFFS* method. This advantage in performance is obtained using 450 genes. *SVMrfe* and *SFFS* performed best when fewer genes/features are retained. In our experiments, we choose the best kernel and the best parameters for each dataset using 10-fold cross validation on the training data.

In the second experiment, we compare several feature transform methods using the stand-alone SVM as the classifier. To reduce the computation time, 1000 genes are first selected by Fisher and then PCA is used to decorrelate the data. In Table 3 the average accuracy on all the datasets obtained using different feature transform methods is reported as a function of the dimension *k* of the projection space ($k \in \{20,30,45\}$) (the accuracy obtained in each dataset is reported in Supplementary Table S2 in the Supplementary Material). The best performance is obtained by *NPE* that only slightly improves the performance obtained by *Fi* in the previous test reported in Table 2.

In the third experiment, we evaluate the performance obtained by varying the image descriptors used to represent the microarray patterns (as described in Section 2.4). In Table 4 we report the accuracy obtained: (i) by methods based on different descriptors; (ii) by the tree wavelet (*TW*) approach (where the classifiers are combined by vote rule); (iii) by the ensemble *FUS* (which is the fusion by vote rule of *TW*, *NPE* and *Fi*) and, as a reference; (iv) by the best approaches previously tested (*Fi* and *NPE*). It is interesting to note in Table 4 that not only does the fusion approach obtain the best average performance but also *FUS* closely matches the performance of the best approach for any given dataset:

- In the prostate dataset (P), the best single approach is *TW*, which *FUS* matches;
- In the breast dataset (B), *NPE* outperforms *TW* and *F*. *FUS* obtains a performance only slightly lower than *NPE* but higher than either *Fi* and *TW*;

Table 4. Average accuracy obtained using different feature transform methods in reduced spaces of different dimensionality *k*

ACC	Fi	NPE	TW	Lu	Lr	LP(3)	LP(5)	LQPu	*FUS*
O	**100.00**	**100.00**	**100.00**	96.40	87.20	91.20	94.00	94.80	**100.00**
P	93.85	95.38	**96.15**	80.00	70.77	65.38	66.15	84.62	**96.15**
L	**100.00**	**100.00**	**100.00**	95.56	93.89	92.22	82.22	98.33	**100.00**
B	82.86	**90.00**	87.14	71.43	74.29	61.43	54.29	84.29	88.57
M	**70.00**	**70.00**	**70.00**	68.33	68.33	68.33	68.33	68.33	66.67
C	**75.00**	68.33	**75.00**	65.00	65.00	65.00	65.00	65.00	73.33
D	87.50	90.00	85.0	72.50	65.00	45.00	45.00	80.00	**90.00**
A	98.57	97.14	95.71	88.57	72.86	65.71	65.71	82.86	**100.00**
DL	**98.57**	**98.57**	**98.57**	75.71	75.71	68.57	68.57	77.14	**98.57**
avg	89.59	89.93	89.73	79.27	74.78	69.20	67.69	81.70	**90.37**

The bold values are the highest performance, the italic values are the values of parameters.

Table 5. Comparison among FUS and different state of the art methods

ACC	*FUS*	LI	CN	GH	LU	PA	HU	BO	CH	OR	PO
O	**100**						100	100			
P	**96.15**			90.16		76.50	96.00		95.09		
L	**100**		99.33	96.38		97.30	99.30	98.89			99.33
B	88.57					73.70					
M	66.67										
C	73.33			82.77	80.72			80.95		85.60	90.00
D	**90.00**				86.83						
A	**100**	100	100	94.52	97.21	100	100	94.46	98.61	94.40	100
DL	98.57				95.56					**98.70**	

The bold values are the highest performance, the italic values are the values of parameters.

- In the ALML dataset (A), *Fi* outperforms *TW* and *NPE*. *FUS*, however, outperforms *Fi*.

We tried combining *LQPr* in *FUS*, but performance remained the same. The most advanced methods based on image descriptors (i.e. *LQPr* and *LQPu*) perform much better than do simple *Lu*, *Lr* and *LP* (we believe, however, that combinations of different texture descriptors with the simple methods would probably obtain performances closer to those obtained by standard approaches).

In the fourth experiment, we compare the performance of *FUS* with several state-of-the art approaches: LI (Liu *et al*., 2002), CN (Cheng, 2010), GH (Ghorai *et al*., 2011), LU (Luo *et al*., 2011), PA (Paliwal and Sharma, 2010), HU (Huerta *et al*., 2010), BO (Bolón-Canedo *et al*., 2012), CH (Chen and Lin, 2011), OR (Orsenigo and Vercellis, 2011) and PO (Porto-Díaz *et al*., 2011).

This comparison shows the goodness of the proposed approach with respect to the state of the art. The only dataset where our results are lower is with the Colon dataset (C). In several of the papers used in Table 5, the feature selection was performed using the training data, but system performance was measured with the testing set, where varying numbers of the features were retained (see Table 9 for the performance of PO using the original code tested in our datasets). In Table 5, we give the best results reported for each method using the testing set. Our method, in contrast, used the same number of

Table 6. Average AUC obtained using different feature selection methods as a function of the number g genes retained

avg(AUC)		Fi	Gi	Mr	Sb	Tt	FFacsa2	SVMrfe	SFFS
g	***150***	89.70	79.17	86.62	85.22	84.01	89.51	89.67	89.70
	300	89.62	86.56	89.21	85.22	85.58	**90.63**	90.06	89.62
	450	90.30	87.20	88.63	85.22	86.34	90.40	90.17	90.30

The bold values are the highest performance, the italic values are the values of parameters.

Table 7. Average AUC obtained using different feature transform methods in reduced spaces of different dimensionality k

avg(AUC)		LLE	OLDA	ONPP	NPE
k	***20***	89.49	89.53	90.06	91.79
	30	89.80	89.53	91.01	91.83
	45	90.33	89.53	91.69	**91.85**

The bold values are the highest performance, the italic values are the values of parameters.

Table 8. AUC obtained by different texture descriptors, *TW*, *Fi*, *NPE* and the ensembles *FUS* and *WF*

AUC	Fi	NPE	TW	Lu	Lr	LP(3)	LP(5)	LQPu	*FUS*	*WF*
O	99.97	99.97	**99.97**	99.72	97.89	99.89	99.89	99.55	**99.97**	**99.97**
P	95.44	96.50	**98.24**	87.28	85.31	89.47	90.45	86.52	97.50	97.71
L	99.97	99.97	**99.97**	99.46	99.46	99.63	**99.97**	98.55	**99.97**	**99.97**
B	94.53	**97.99**	91.08	89.93	80.22	88.45	90.58	91.74	97.11	97.33
M	61.17	**69.13**	66.62	49.04	43.00	46.73	49.55	45.44	66.55	66.82
C	68.19	65.51	**72.10**	57.69	54.52	59.89	66.00	45.42	69.02	69.17
D	93.61	97.95	98.21	81.33	64.71	79.54	89.77	82.86	98.66	**98.72**
A	99.95	99.95	**99.95**	**99.95**	93.98	99.23	99.77	99.05	**99.95**	**99.95**
DL	99.95	99.76	**99.95**	94.89	94.08	95.98	98.82	94.70	**99.95**	**99.95**
avg	90.30	91.85	91.78	84.36	79.24	84.31	87.20	82.64	92.07	**92.18**

The bold values are the highest performance, the italic values are the values of parameters.

features both in training and testing as well as across all datasets. Our method is thus very suitable for general practitioners.

In Tables 6–9, we report results obtained in the previous experiments using a more reliable performance indicator: the AUC. AUC can be interpreted as the probability that the classifier will assign a lower score to a randomly picked positive sample than to a randomly picked negative sample.

In Table 6, we compare several feature selection methods using AUC (cf. Table 2 where we used accuracy as the performance indicator). *FFacsa2* provides the best performance. It should be noted that this difference is mainly due to the lower performance obtained by the other methods in the **M** dataset (see Supplementary Table S3 in the Supplementary Material for results of each dataset).

In Table 7, we compare the different feature transform techniques using AUC. The best performance, as in Table 3 using accuracy, is obtained by *NPE*.

In Table 8, we report the performance obtained in the third experiment. In this Table a new ensemble is evaluated, WF, which is the fusion by weighted sum rule of *TW*, *NPE*, *Fi* and *LP(5)*.

Table 9. Comparison of *WF* with different state of the art methods using AUC as the performance indicator

AUC	LIU2	OldTW	K = 64	K = 128	K = 512	K = 50%	OC	WF
O	99.97	99.97	99.97	99.97	99.97	**99.97**	99.30	**99.97**
P	97.00	96.70	94.61	95.82	95.14	96.42	93.47	**97.71**
L	99.90	**99.97**	**99.97**	**99.97**	**99.97**	**99.97**	**99.97**	**99.97**
B	86.50	91.10	92.60	94.70	95.52	96.83	93.38	**97.33**
M	–	–	58.60	54.81	61.94	52.95	61.75	**66.82**
C	–	–	**69.41**	68.44	68.44	69.17	63.37	69.17
D	–	–	97.70	98.21	89.77	89.77	95.91	**98.72**
A	–	–	**99.95**	**99.95**	**99.95**	**99.95**	**99.95**	**99.95**
DL	–	–	99.38	**99.95**	99.76	99.76	98.25	**99.95**
avg	–	–	90.24	90.20	90.05	89.42	89.48	**92.18**

The bold values are the highest performance, the italic values are the values of parameters.

In the weighted sum rule, each classifier is weighted by a value between 0 and 1. The scores are then summed. Optimal weights are obtained using the leave-one-dataset-out validation protocol. In other words, when a given dataset is classified, the sets of weights are selected using as the validation set the others datasets. Our fusion approach WF obtains the best overall average performance using AUC. Moreover, fusion results for each dataset closely approximate the performance of the best methods reported for the individual datasets.

In Table 9, we compare our best approach *WF* with the performance obtained by a random subspace of SVM trained using the original genes, LIU2 (Liu, 2009), and OldTW (Nanni and Lumini, 2011). Random subspace of SVM has been shown to be very effective (Bertoni *et al*., 2009). The random subspace creates an ensemble such that each classifier is trained with a different subset of the original features. In our experiments, we combine results with sum rule using 50 classifiers, each trained with K features. In Table 9, $K = 50\%$ means that each classifier is trained with a subset that contains 50% of the original features, whereas $K = x$ means that each classifiers is trained with x randomly selected genes. **PO** in Table 9 refers to the results obtained using the original code shared by (Porto-Díaz *et al*., 2011) with the following setting: we ran their approach starting from the 500 genes selected by *Fi* (in this way a more fair approach with our method is provided). It is interesting to note that now the performance on the Colon dataset (C) is lower than that obtained by our ensemble. *WF* outperforms the other methods.

The advantage of using a combination of approaches is also demonstrated by the use of the Wilcoxon Signed-Rank test (Demsar, 2006) developed for comparing the results of stand-alone methods with ensembles. The null hypothesis (that is there is no difference between the accuracies of the stand-alone methods and the ensemble) is rejected with a level of significance of 0.10.

As an additional experiment, we investigated the relationship among the different approaches by evaluating the error independence between the classifiers trained using those features. Table 10 reports the average Yule's Q-statistic (Kuncheva and Whitaker, 2003) in the tested datasets. For two classifier G_i and G_j the Q-statistic, a posteriori measure, is defined as:

where N^{ab} is the number of instances in the test set, classified correctly ($a = 1$) or incorrectly ($a = 0$) by the classifier G_i, and correctly ($b = 1$) or incorrectly ($b =$) by the classifier G_j. $Q \in [-1, 1]$

Table 10. Yule's Q-statistic between the stand-alone approaches

compared descriptors	O	P	L	B	M	C	D	A	DL
FI versus NP	1.00	0.96	0.93	0.98	0.99	0.60	1.00	0.99	1.00
FI versus TW	1.00	0.99	0.93	0.94	0.93	0.63	0.98	1.00	1.00
NPE versus T	1.00	0.96	1.00	0.93	0.95	0.99	0.99	0.99	1.00

Table 11. AUC obtained in the datasets used in (Shi *et al*., 2011)

compared descriptors	*K*= 64	K = 128	K = 512	FFacsa2	FV	Fi	NPE	TW	WF
WB	76.41	76.41	75.72	64.58	69.56	73.50	67.93	72.97	**74.19**
LA	65.12	68.54	69.03	65.76	66.27	71.53	**73.02**	67.04	71.31
avg	70.76	72.47	72.37	65.17	67.91	72.51	70.48	70.00	**72.75**

The bold values are the highest performance, the italic values are the values of parameters.

and $Q_{i,j} = 0$ for statistically independent classifiers. Classifiers that tend to recognize the same patterns correctly will have $Q > 0$, and those that commit errors on different patterns will have $Q < 0$. In this problem, the Q-statistic values are low enough to validate the idea of combining the different approaches.

As a final experiment, in Table 11, we report the results of our ensemble on two other recent datasets from (Shi *et al*., 2011). The first is a breast cancer dataset (WB) that contains a subset of ER-positive, lymphnode-negative patients who did not received adjuvant treatment. The raw intensity Affymetrix CEL files and normalized data by RMA procedures using Bioconductor packages are used for obtaining a final expression matrix comprising 22 283 features and 209 samples. The 71 patients who developed distant metastases or died within 5 years are classified as poor prognosis subjects, and the 139 patients who remained healthy for >5 years are classified as good prognosis subjects. The second dataset (LA) contains gene expressions of 86 patients with primary lung ADCA; 62 patients were still alive, and 24 patients had died.

Notice that all the parameters of WF are obtained using the nine datasets previously used throughout this article. In this test, we arrive at the same main conclusion of the previous test: the fusion, WF, obtains the best average performance.

4 CONCLUSION

The goal of this study was to develop a robust ensemble of SVM classifiers based on feature perturbation for microarray classification. The reported results of our experiments, expressed as both accuracy and AUC, show that our approach performs very well across several datasets. Our study examined an SVM trained using a set of selected genes by Fisher criterion, an SVM trained using the feature set obtained by NPE, a set of SVMs trained using a set of orthogonal wavelet coefficients of different wavelet mothers and a set of SVMs trained using texture descriptors extracted from the microarray, considering it as an image. The positive results we obtain compare well with those reported in the literature and provide further confirmation that ensembles of classifiers obtain more reliable results.

In future studies, we plan on testing our approach using more datasets. We will also study combining additional methods in ensemble construction (e.g. combining our feature perturbation approaches with a pattern perturbation approach).

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. Natl Acad. Sci. USA*, **96**, 6745–6750.

Beer,D.G. *et al.* (2002) Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat. Med.*, **8**, 816–823.

Ben-Dor,A. *et al.* (2003) Tissue classification with gene expression profiles. *J. Comput. Biol.*, **7**, 559–583.

Bertoni,A. *et al*. (2009) Classification of DNA microarray data with random projection ensembles of polynomial. In *18th Italian Workshop on Neural Networks*. IOS Press, Vietri sul Mare, Italy. pp. 60–66.

Bolón-Canedo,V. *et al.* (2012) An ensemble of filters and classifiers for microarray data classification, *Pattern Recognit.*, **45**, 531–539.

Bontempi,G. (2007) A blocking strategy to improve gene selection for classification of gene expression data. *IEEE/ACM Trans. Comput. Biol. Biofrom.*, **4**, 293–300.

Broet,P. *et al.* (2004) A mixture model-based strategy for selecting sets of genes in multiclass response microarray experiments. *Bioinformatics*, **20**, 2562–2571.

Brown,M.P. *et al.* (2000) Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl Acad. Sci. USA*, **97**, 262–267.

Chee,M. *et al.* (1996) Assessing genetic information with high-density dna arrays. *Science*, **274**, 610–614.

Chen,A.H. and Lin,C.-H. (2011) A novel support vector sampling technique to improve classification accuracy and to identify key genes of leukaemia and prostrate cancer *Expert Syst Appl*, **38**, 3209–3219.

Cheng,Q. (2010) A sparse learning machine for high-dimensional data with application to microarray gene analysis. *IEEE/ACM Trans. Comput. Biol. Biofrom.*, **7**, 636–646.

Demsar,J. (2006) Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.*, **7,** 1–30.

Der,S.D. *et al.* (1998) Identification of genes differently regulated by interferon alpha, beta, or gamma using oligonucleotide arrays,. *Proc. Natl Acad. Sci. USA*, **95**, 15623–15628.

Devore,J. and Peck,R. (1997) *Statistics: the Exploration and Analysis of Data*. Duxbury Press, Florence, KY.

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

Figuiredo,M.A.T. and Jain,A.K. (2001) Baysean learning of sparse classifiers. In *Computer Vision and Pattern Recognition (CVPR '01)*. IEEE Computer Society, Miami, Florida, pp. I-35-I-45.

Freije,W.A., *et al.* (2004) Gene expression profiling of gliomas strongly predicts survival. *Cancer Res.*, **64**, 6503–6510.

Fung,G. and Mangasarian,O.L. (2000) Data selection for support vector machine classifiers. In *Association for Computing Machinery Special Interest Group on Knowledge Discovery and Data Mining*. Association for Computing Machinery, New York, USA, pp. 64–70.

Furey,T.S. *et al.* (2000) Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*. **16**, 906–914.

Ghorai,S. *et al.* (2011) Cancer classification from gene expression data by NPPC ensemble. *IEEE/ACM Trans. Comput. Biol. Biofrom.*, **8**, 659–671.

Ghosh,D. and Chinnaiyan,A.M. (2005) Classification and selection of biomarkers in genomic data using LASSO. *J. Biomed. Biotechnol.*, **2**, 147–154.

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

Gordon,G.J. *et al.* (2002) Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. *Cancer Res.*, **62**, 4963–4967.

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

Hastie,T. *et al*. (2009) *The Elements of Statistical Learning*. Springer, New York.

Hayat,M. and Khan,A. (2011) Predicting membrane protein types by fusing composite protein sequence features into pseudo amino acid composition. *J. Theor. Biol.*, **271**, 10–17.

He,X. *et al.* (2005) Neighborhood preserving embedding. *Tenth IEEE International Conference on Computer Vision (ICCV'2005)*, IEEE Computer Society, Beijing, China.

Huang,J. *et al.* (2010) Decision forest for clssification of gene expression data. *Comput. Biol. Med.*, **40**, 698–704.

Huang,T.M. and Keoman,V. (2005) Gene extraction for cancer diagnosis by support vector machines-an improvement. *Artif. Intel. Med.*, **40**, 185–194.

Huerta,E.B. *et al.* (2010) A hybrid LDA and genetic algorithm for gene selection and classification of microarray data. *Neurocomputing*, **73**, 2375–2383.

Kokiopoulou,E. and Saad,Y. (2005) Orthogonal Neighborhood Preserving Projections. *IEEE International conference on Data Mining*. IEEE Computer Society, New Orleans, LA.

Kudo,M. and Sklansky,J. (2000) Comparison of algorithms that select features for pattern classifiers. *Pattern Recognit.*, **33**, 25–41.

Kuncheva,L.I. and Whitaker,C.J. (2003) Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Mach. Learn.*, **51**, 181–207.

Lai,Y. *et al.* (2004) Statistical method for identifying diferential gene-gene coexpression patterns. *Bioinformatics*, **20**, 3146–3155.

Lee,G. *et al.* (2008) Investigating the efficiacy of nonlinear dimensionality reduction schemes in classifying gene- and protein-expression studies. *IEEE/ACM Trans. Comput. Biol. Biofrom.*, **5**, 368–384.

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,Y. (2009) Wavelet feature extraction for high dimensional microarray data. *Neurocomputing*, **72**, 985–990.

Luo,L. *et al.* (2011) Methods of forward feature selection based on the aggregation of classifiers generated by single attribute. *Comput Biol Med.*, **41**, 435–441.

Maglietta,R. *et al.* (2007) Selection of relevant genes in cancer diagnosis based on their prediction accuracy. *Artif. Intel. Med.*, **40**, 29–44.

Nanni,L. and Lumini,A. (2011) Wavelet selection for disease classification by DNA microarray data. *Expert Syst Appl.*, **38**, 990–995.

Nanni,L. *et al.* (2010) Local binary patterns variants as texture descriptors for medical image analysis. *Artif. Intel. Med.*, **49**, 117–125.

Ojala,T. *et al.* (2002) Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Mach. Intell.*, **24**, 971–987.

Ojansivu,V. and Heikkila,J. (2008) Blur insensitive texture classification using local phase quantization. In *International Conference on Image and Signal Processing* . Springer, Cherbourg-Octeville, France, pp. 236–243.

Orsenigo,C. and Vercellis,C. (2011) An effective double-bounded tree-connected isomap algorithm for microarray data classification. *Pattern Recognit. Lett.*, **33**, 9–16.

Paliwal,K.K. and Sharma,A. (2010) Improved direct LDA and its application to DNA microarray gene expression data. *Pattern Recognit. Lett.*, **31**, 2489–2492.

Peng,H. *et al.* (2005) Feature selection on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intel.*, **27**, 1226–1238.

Peng,Y. (2006) A novel ensemble machine learning for robust microarray data classification. *Comput. Biol. Med.*, **36**, 553–573.

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

Pomeroy,S.L. *et al.* (2002) Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature*, **415**, 436–442.

Porto-Díaz,I. *et al.* (2011) A study of performance on microarray data sets for a classifier based on information theoretic learning. *Neural Netw.*, **24**, 888–896.

Pudil,P. *et al.* (1994) Floating search methods in feature selection. *Pattern Recognit. Lett.*, **5**, 1119–1125.

Roth,V. (2004) The generalized LASSO. *IEEE Trans. Neural Netw.*, **15**, 16–18.

Roweis,S. and Saul,L. (2000) Nonlinear dimensionality reduction by locally linear embedding. *Science*, **290**, 2323–2326.

Shen,L. and Tan,E.C. (2005) Dimension reduction-based penalized logistic regression for cancer classification using microarray data. *IEEE/ACM Trans. Comput. Biol. Biofrom.*, **2**, 166–175.

Shi,P. *et al.* (2011) Top scoring pairs for feature selection in machine learning and applications to cancer outcome prediction. *BMC Bioinformatics*, **12**, 375.

Shipp,M.A. *et al.* (2002) Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nat. Med.*, **8**, 68–74.

Singh,D. *et al.* (2002) Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, **1**, 203–209.

Statnikov,A. *et al.* (2008) A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. *BMC Bioinformatics*, **9**, 319.

Stiglic,G. *et al.* (2010) Finding optimal classifiers for small feature sets in genomics and protoemics. *Neurocomputing*, **73**, 2346—2352

Tahir,M. *et al.* (2011) Protein subcellular localization of fluorescence imagery using spatial and transform domain features. *Bioinformatics*. doi:10.1093/bioinformatics/btr624

Tamayo,P. *et al.* (1999) Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl Acad. Sci. USA*, **96**, 2907–2912.

Tan,A.C. and Gilbert,D. (2003) Ensemble machine learning on gene expression data for cancer classification. *Appl. Bioinformatics*, **2**, 75–83.

Tibshirani,R. *et al.* (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl Acad. Sci. USA*, **99**, 6567–6572.

Tibshirani,R. *et al.* (2003) Class predition by nearest shrunken centroids, with application to DNA microarrays. *Stat. Sci.*, **18**, 104–117.

Turashvilli,G. *et al.* (2007) Novel markers for differentiation of lobular and ductal invasive breast carcinomas by laser microdissection and microarry analysis. *BMC Cancer*, **7**, 55.

van 't Veer,L.J. *et al.* (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**, 530–536.

Ye,J. (2005) Characterization of a family of algorithms for generalized discriminant analysis on undersampled problems. *J. Mach. Learn. Res.*, **6**, 483–502.