THE INTERNATIONAL JOURNAL OF SCIENCE & TECHNOLEDGE

Microarray Data Analysis Using Masked Sequential Backward Selection

V. Devi
Research Scholar, Gurunanak College, Chennai, India
S. Natarajan
Department of Plant Biology and Biotechnology,
Gill Research Institute, Guru Nanak College, Chennai, India

Abstract:
In any microarray data analysis, selecting the optimal number of features is a very important task without information loss and removing unwanted data. Also, the coverage of training samples should be high considering only the best features in a dataset. In this paper, a new method called masked sequential backward selection technique is proposed and its performance is compared with other popular feature selection techniques. The method selects the most differentially expressed genes based on its expression value distribution. Also, it allows the user to select the number of features to be retrieved to get best classification accuracy and study a larger set of high relevant genes for the target disease.

Keywords: *Microarray, feature selection, expression, classification*

1. Background

The most difficult task is identifying the optimal number of features, as it should balance between including important data and removing unwanted information. The minimum number of genes selection discussed in previous chapter has improved data classification accuracy on the training data. The next step in microarray data analysis is to select the differently expressed genes based on expression values. This chapter presents a new filter feature selection that selects the discriminative genes to reach the best classification accuracy. This method allows the user to define number of features to reach best classification accuracy and to retrieve a large number of relevant genes for the target disease. This method is named masked sequential backward selection as it combines the features of Masked Painter feature selection method and Sequential backward selection algorithm. It adopts the sequential feature selection and has the following two important features:

- It finds the minimum number of genes with best coverage of training data and maximizes the correct assignment of the training samples to the corresponding class.
- It ranks the genes according to a quality score which is computed to remove the effect of outliers.

The top ranked genes along with the minimum gene subset are pooled to obtain a better classification accuracy. The algorithm is based on identification of genes that are overlapping in their expression and the information in the gene are hidden and stored as gene masks. Sequential feature selection starts with an empty feature subset which iteratively adds features until user defined size is reached. This method was validated on different multi-category microarray datasets to give a more practical evaluation of the proposed method. Also, the masked sequential backward selection was compared with other feature selection methods to determine its classification accuracy. In most of the cases the proposed method gives higher accuracy statistically. Finally, the biological relevance of the selected genes of the proposed approach was evaluated.

2. Method

The principle behind masked sequential backward selection is that few genes can be easily identified as from a sample belonging to a particular class since their expression level does not overlap with other classes. But, in other cases, it is really difficult to differentiate between classes as they have overlapping ranges. To discriminate such genes, the proposed method of masked sequential backward selection is applied. This method primarily represents each gene by a gene mask based on its ability to assign training samples to the correct class. Then, it calculates overlap score and dominant class which are pooled together for the ranking of genes. The overlap score is an index of the overlap in the expression intervals for different classes. In contrast, dominant class of a gene is defined by the class to which majority of the samples belong without overlap in their expression. Two different scoring systems are employed since genes with less overlapping intervals create ambiguity when assigning correct class. It is sequentially iterated until user defined size of the genes in a class is obtained. The method has the following phases:

2.1. Gene Mask Computations

The method starts by assigning a mask to every gene, denoted by an array with number of samples ('S' bits). Masking is done to represent the gene's capability to classify each sample. The mask is created for a gene 'i','s' its mask will be assigned 1 if its expression value 'e_{is}' falls within the expression interval of a single class, or else 0.

Formally, given two arbitrary classes c1, c2 ∈ C = {1. . . k}, bit s of gene mask i is computed as follows.
$mask_{is}$ =

$$\begin{cases} 1 \text{ if } (e_{is} \in Ii,c1) \mid e_{is} \in Ii,c2 \\ 0 \text{ otherwise} \end{cases} \quad \text{---------- 1}$$

If the sample did not fall under any core expression interval, it is an outlier and its value is set to 0.

2.2. Overall Score Computation and Dominant Class Assignment

After gene mask computation, each gene is assigned with an overlap score which is the degree of overlap among the core expression intervals. The core expression intervals are computed using Weighted Mean Deviation (WMD) method followed by assigning dominant class for each gene. Dominant class is the best distinguished class and it is done to reduce redundancy in gene selection.

Based on the degree of overlap in the expression intervals among classes, overlap score is assigned to each gene. Overlap score differs from gene mask score as gene max is the min-max expression interval whereas overlap score is the representation of the differentiating degree of genes and its aim is to avoid over fitting through noise and outliers handling. Thus, to model the probable values of a unknown test set, the overlap score is based on core expression intervals which is calculated using Weighted Mean deviation (WMD) method.

Overlapping intervals in minimum gene subset will lead to wrong classifications, because the differentiating power of the gene is inadequate. Hence, for ranking genes the overlap score is used which will be higher for less important genes which has many overlaps among the different classes. The lower the overlap score represents genes with high differentiating power with less overlap among their expression intervals. Each gene is assigned an overlap score, represented as 'os', is based on the following features in the gene expression values.

- The number of samples linked to different classes of the same range.
- The number of overlapping classes
- The overlapping interval length

The overlap score is represented as 'osi' for each gene 'i'. The total expression interval represented by 'W' of a gene (amplitude=|W|) is calculated by range, which is the difference in the minimum and maximum values among its core expression interval boundaries. The total expression interval represented as 'W' of a gene is calculated by finding the range of its core expression interval boundaries (subtracting the minimum and maximum of the core expression interval boundaries). For each total expression interval, subintervals are created based on the different set of overlapping classes with respect to the adjacent subintervals. For example, the subinterval wt can be defined as the interval enclosed by two successive extremes of core expression intervals and its amplitude can be represented as |wt|. The genes which have more overlaps among expression intervals of different class will get higher scores which is based on both on number of samples of different classes with similar subinterval and the amplitude of the subinterval. Based on this, the overlap score for any gene can be defined as follows:

$$os = \sum_{t=1}^{T} c_t \cdot m_t / M \cdot |wt|/|W|$$

whereT is the number of subintervals, ct is the number of classes which overlap in subinterval t, mt is the number of samples expressed in subinterval t, and M is the total number of samples. Subintervals covered by a single class provide no contribution to the overlap score, because the number of overlapping classes is 0. In the case of subintervals without values, the number of overlapping classes is 0. Thus, also in this case, no contribution is added to the overlap score.

A dominant class of a gene is the class to which it differentiates best. To identify the dominant class of a gene, the subintervals to which the expressed samples of a simple class is considered and the number of samples in that class is identified. The class containing the highest number of samples is the dominant class of the gene as this will improve the classification accuracy.

2.3. Minimum Gene Subset Selection

The best training set sample coverage can be outfitted by selecting the minimum number of genes using overlaps scores and searching algorithms such as K-means clustering with greedy and set covering approaches.

2.4. Gene Ranking

The genes which are not part of the minimum subset were ranked according to the higher overlap scores, separately for each dominant class. The final gene rank was created selecting the top ranked genes from each dominant class.

Next, the genes are ranked based on overlap score and dominant class. All the genes, except those belonging to a minimum gene subset are ranked based on increasing value of overlap score for each dominant class. The topmost gene from each dominant class was selected and the final rank was prepared.

2.5. Final Gene Selection

The top ranked genes were selected by Sequential backward selection algorithm and supplemented with the minimum gene subset to create the final gene set.

To provide best sample coverage on the training set, minimum gene subset was prepared which is extended to include top 'k' ranked genes using sequential backward selection algorithm. This approach will include maximum set of training samples as the minimum gene subset is independent of overlap score.

3. Results and Discussion

3.1. Microarray Dataset

The masked sequential backward selection was applied on the publicly available datasets on [1], [2], and [3]. Seven multi-category microarray datasets were chosen, where five among them are characterized by 3 to 9 classes and the rest two contain only two classes. The number of features in each class ranges from 2,000 to more than 10,000. The characteristics of the dataset are presented in Table 1.

Dataset	Genes	Samples	Classes
Alon	62	2000	2
Brain1	90	5920	5
Brain2	50	10367	4
Leukemia	72	5327	3
Srbct	83	2308	4
Tumour9	60	5727	9
Welsh	34	7129	2

Table 1: Characteristics of Dataset

The masked sequential backward selection (SB) method was authenticated by comparing it with other feature selection techniques with gene expression datasets. The feature selection techniques taken for comparison in the present study are Information Gain (IG), Twoing Rule (TR), Sum Minority (SM), Max Minority (MM), Gini Index (GI), Sum of Variance (SV). As these methods are considered standards for microarray datasets [4,5] they are used for comparing with sequential backward selection technique also. The parameters according to [6,7,8] such as number of features, feature selection algorithm, classifier and dataset were cross validating through 50 iterations. The mean classification accuracy on the 50 different iterations was calculated. The significance of the results were tested using Student's 't' test for each iteration. The statistically significant values at 5% level of significance (p-value$\leq$0.05) are indicated using '*' and the best values are represented in bold.

3.2. The Characteristics and the Experimental Results are Presented Below

Classification Accuracy: Classification accuracy is defined by the number of samples correctly classified including true positives and true negatives. This measure is compared for masked sequential backward selection (SB) and other feature selection techniques. Classification accuracy has been tested on all the datasets in Table 1 with J48 decision tree classifier [9]. The results obtained for Alon dataset is presented in Table 2.

#	SB	IG	TR	SM	MM	GI	SV
2	**78.94**	74.20*	74.68*	74.84*	75.81*	74.68*	74.68*
4	**76.95**	73.88*	73.24*	74.01*	75.23*	73.24*	73.24*
6	**7.37**	73.73*	73.75*	74.01*	75.11*	73.75*	73.75*
8	**7.05**	73.79*	73.68*	74.38*	74.83*	73.68*	73.68*
10	**76.85**	73.94*	73.77*	74.15*	74.99*	73.77*	73.77*
12	**6.02**	74.07*	73.99*	74.22*	74.12*	73.99*	73.99*
14	**76.15**	73.39*	73.80*	74.24*	74.31*	73.80*	73.80*
16	**75.26**	73.07*	73.19*	73.75*	74.11	73.19*	73.19*
18	**5.65**	73.00*	73.05*	73.50*	75.23	73.05*	73.05*
20	**75.63**	73.13*	73.08*	73.25*	75.29	73.08*	73.08*
avg	**76.59**	73.62*	73.63*	74.03*	74.90*	73.63*	73.63*
max	**78.94**	74.20	74.68	74.84	75.81	74.68	74.68
SD	1.03	0.42	0.48	0.43	0.53	0.48	0.48

Table 2: Accuracy yielded by the J48 classifier on the Alon dataset.

In Table 2, each row presents the accuracy of a specific cardinality of the different feature selection technique for the Alon dataset. The average accuracy, maximum and standard deviation for each method are presented at the end of the table.

#	SB	IG	TR	SM	MM	GI	SV
2	**82.72**	81.67	80.00*	77.83*	78.25*	81.89	82.47
4	**86.50**	84.36*	81.75*	83.72*	82.50*	84.44*	85.78
6	**86.69**	85.17*	84.06*	85.44*	83.81*	85.42*	85.53
8	**86.44**	85.53	85.25	85.53	3.78*	86.14	85.06*
10	**86.86**	85.39*	85.22*	85.89	4.42*	85.75*	84.94*
12	**86.83**	85.14*	85.14*	85.69*	85.56*	85.56*	85.11*
14	**86.72**	84.97*	84.92*	85.11*	85.42*	85.25*	85.50*
16	**86.58**	84.92*	84.89*	85.11*	85.28*	85.11*	85.69
18	**86.67**	84.69*	84.72*	84.97*	85.03*	84.94*	86.17
20	**87.22**	84.86*	84.86*	85.03*	85.36*	84.97*	86.44
avg	**86.32**	84.67*	84.08*	84.43*	83.94*	84.95*	85.27*
max	**87.22**	85.53	85.25	85.89	85.56	86.14	86.44
SD	1.21	1.05	1.68	2.27	2.11	1.11	1.04

Table 3: Accuracy yielded by the J48 classifier on the Leukemia dataset

Table 3, presents the accuracy of a specific cardinality of the different feature selection technique for the Leukemia dataset. The average accuracy, maximum and standard deviation for each method are presented at the end of the table.

#	SB	IG	TR	SM	MM	GI	SV
2	**71.50**	65.37*	63.76*	59.51*	62.41*	65.79*	63.63*
4	**81.73**	75.60*	72.97*	69.23*	69.89*	74.57*	74.00*
6	**81.92**	78.18*	75.17*	75.06*	72.72*	75.96*	78.05*
8	**82.07**	78.94*	76.75*	76.63*	75.18*	77.32*	80.61*
10	**82.07**	79.52*	78.06*	78.21*	77.18*	78.29*	81.02
12	**82.09**	80.63*	78.99*	78.68*	79.02*	79.64*	80.93*
14	**81.48**	80.85	79.80*	78.37*	80.75	80.54*	81.23
16	81.07	81.11	80.48	78.10*	81.13	81.20	**82.10**
18	81.16	81.18	81.01	78.23*	81.71	81.51	**82.71***
20	80.61	81.06	81.25	78.28*	82.48*	81.79*	**82.78***
avg	**80.57**	78.24*	76.82*	75.03*	76.25*	77.66*	78.71*
max	82.09	81.18	81.25	78.68	82.48	81.79	**82.78**
dev	3.06	4.60	5.04	5.85	6.06	4.59	5.60

Table 4: Accuracy yielded by the J48 classifier on the Srbct dataset

In Table 4, each row presents the accuracy of a specific cardinality of the different feature selection technique for the Srbct dataset. The average accuracy, maximum and standard deviation for each method are presented at the end of the table. The Sequential Backward Eventually, on the Srbctdataset it is outperformed by the SV technique for larger sets of features (18and 20). However, its overall average performance is statistically better thanall other methods.

#	SB	IG	TR	SM	MM	GI	SV
2	70.27	68.89*	67.62*	**70.65**	68.88*	68.42*	69.01*
4	**72.19**	70.45*	69.99*	71.36	69.93*	69.72*	71.01
6	**73.22**	71.41*	70.76*	71.25*	70.50*	70.84*	71.25*
8	**73.11**	72.35	71.09*	1.13*	70.95*	70.78*	72.43
10	**73.25**	72.38	71.48*	71.16*	71.25*	71.27*	72.61
12	**73.25**	72.79	72.06	1.64*	71.97	71.77*	72.74
14	**73.39**	73.19	72.55	72.16	71.76*	71.60*	72.97
16	**73.54**	73.46	72.95	72.77	71.70*	72.22	73.25
18	**74.11**	73.71	73.58	73.03	72.20*	72.79	73.42
20	**74.49**	73.80	73.65	73.22*	71.72*	73.05*	73.37
avg	**73.08**	72.24*	71.57*	71.84*	71.09*	71.25*	72.21*
max	**74.49**	73.80	73.65	73.22	72.20	73.05	73.42
dev	1.10	1.50	1.74	0.85	0.99	1.33	1.32

Table 5: Accuracy yielded by the J48 classifier on the brain1 dataset

Table 5, presents the accuracy of a specific cardinality of the different feature selection technique for the brain1 dataset. The average accuracy, maximum and standard deviation for each method are presented at the end of the table.

#	SB	IG	TR	SM	MM	GI	SV
2	**57.64**	46.80*	46.80*	46.13*	49.03*	49.35*	46.80*
4	**58.23**	46.11*	46.11*	46.15*	51.78*	49.38*	46.11*
6	**58.83**	45.84*	45.84*	48.45*	53.12*	49.01*	45.84*
8	**59.19**	46.77*	46.77*	49.02*	54.66*	49.33*	46.77*
10	**59.43**	48.16*	48.16*	51.29*	55.51*	50.29*	48.16*
12	**59.27**	49.65*	49.53*	54.30*	56.41*	51.11*	49.65*
14	**59.73**	50.00*	49.96*	55.87*	56.61*	51.50*	49.92*
16	**60.04**	50.58*	50.26*	56.77*	57.15*	52.73*	50.33*
18	**59.85**	50.94*	51.20*	56.96*	57.33*	53.48*	50.58*
20	**59.62**	51.27*	50.90*	57.24*	57.75	54.23*	50.91*
avg	**59.18**	48.61*	48.55*	52.22*	54.93*	51.04*	48.51*
max	**60.04**	51.27	51.20	57.24	57.75	54.23	50.91
dev	0.72	2.00	1.95	4.31	2.68	1.80	1.89

Table 6: Accuracy yielded by the J48 classifier on the brain2 dataset

Table 6, each row presents the accuracy of a specific cardinality of the different feature selection technique for the brain2 dataset. The average accuracy, maximum and standard deviation for each method are presented at the end of the table.

#	SB	IG	TR	SM	MM	GI	SV
2	**25.40**	24.30	23.03*	21.47*	18.63*	21.13*	24.77
4	29.33	28.77	28.30	24.00*	20.80*	23.43*	**29.77**
6	30.97	30.43	30.47	25.83*	20.73*	24.33*	**31.27**
8	**31.03**	32.30	31.07	28.00*	21.40*	24.67*	32.17
10	**31.57**	32.77	31.63	28.27*	22.37*	26.60*	31.50
12	**32.03**	32.60	31.40	29.80*	21.87*	26.97*	31.60
14	**31.97**	32.77	30.97	29.77*	21.13*	27.33*	31.80
16	**31.97**	32.83	30.97	28.50*	22.23*	26.90*	31.17
18	**32.07**	32.87	31.23	29.60*	23.27*	27.63*	30.33
20	**33.03**	33.07	30.80*	29.63*	23.43*	27.70*	30.90*
avg	**30.94**	31.27	29.99*	27.49*	21.59*	25.67*	30.53
max	**33.03**	33.07	31.63	29.80	23.43	27.70	32.17
dev	2.06	2.67	2.48	2.70	1.33	2.08	2.03

Table 7: Accuracy yielded by the J48 classifier on the Tumour9 dataset

In Table 7, each row presents the accuracy of a specific cardinality of the different feature selection technique for the Tumour9 dataset. The average accuracy, maximum and standard deviation for each method are presented at the end of the table. On the Tumor9 dataset, the Sequential backward method shows a performance equivalent with the best techniques (IG and SV).

#	SB	IG	TR	SM	MM	GI	SV
2	**89.72**	85.81*	85.81*	85.81*	85.81*	85.81*	85.81*
4	**90.21**	85.74*	85.74*	85.74*	85.74*	85.74*	85.74*
6	**90.03**	84.72*	84.72*	84.72*	84.72*	84.72*	84.72*
8	**90.24**	85.08*	85.08*	85.08*	85.08*	85.08*	85.08*
10	**89.56**	84.26*	84.26*	84.26*	84.26*	84.26*	84.26*
12	**90.10**	83.58*	83.58*	83.58*	83.58*	83.58*	83.58*
14	**89.97**	83.26*	83.26*	83.26*	83.26*	83.26*	83.26*
16	**90.08**	83.30*	83.30*	83.30*	83.30*	83.30*	83.30*
18	**89.56**	83.31*	83.31*	83.31*	83.31*	83.31*	83.31*
20	**89.30**	83.23*	83.23*	83.23*	83.23*	83.23*	83.23*
avg	**89.88**	84.23*	84.23*	84.23*	84.23*	84.23*	84.23*
max	**90.24**	85.81	85.81	85.81	85.81	85.81	85.81
dev	0.30	0.99	0.99	0.99	0.99	0.99	0.99

Table 8: Accuracy yielded by the J48 classifier on the Welsh dataset

Table 8, presents the accuracy of a specific cardinality of the different feature selection technique for the Welsh dataset. The average accuracy, maximum and standard deviation for each method are presented at the end of the table.

From Tables 2 to 8 it is clear that the Masked sequential Backward selection technique is more accurate on all the datasets including Alon, brain1, brain2, leukemia and Welsh. All the results are statistically significant at 99% level as depicted by a '*' in the table values. On an average, Masked sequential backward selection method shows an improvement of +5.65% on all cardinalities of the feature set than the second best method on the Welsh dataset.

3.3. Cardinality of the Selected Feature Set

It is the number of genes present in a particular feature set. The influence on classification accuracy of different numbers of selected gene was analyzed with masked sequential backward selection method and other feature selection methods. The Masked Sequential backward selection method was analyzed varying the cardinality of the selected feature set. The proposed method showed an improvement than the second best method and results are tabulated in Table 9.

Features	Accuracy improvement
2	+3.08%
4	+2.84%
6	+2.81%
8	+1.87%
10	+1.79%
12	+1.91%
14	+1.79%
16	+1.41%
18	+1.11%
20	+1.08%
average	+2.14%

Table 9: Average accuracy improvement over the second best method on all datasets.

It is evident from the Table that the Masked Backward selection method shows higher improvement for small number of selected features. On the other hand, as the cardinality of the selected feature set increases the performance difference decreases for all methods, whereas Masked Backward sequential selection shows a higher accuracy. Thus, the proposed method can perform independently based on the dataset characteristics because its performance is based on the data distribution.

3.4. Minimum Gene Subset Definition

It represents the minimum number of genes present in a particular subset. The performance of masked sequential backward selection method and other feature selection methods were tested.

Dataset	Greedy			Set covering			Max accuracy with fixed #genes	
	#genes	acc.	time [sec.]	#genes	acc.	time [sec.]	genes	acc.
Alon	5.09	71.82%	.137	4.68	70.33%	16.255	2	78.94%
Brain1	6.33	70.23%	1.229	5.59	70.11%	938.556	20	74.49%
Brain2	5.07	57.49%	1.927	4.62	56.52%	35.142	16	60.04%
Leukemia	4.15	87.00%	0.529	3.82	85.89%	46.098	20	87.22%
Srbct	6.51	1.09%	0.246	5.95	79.55%	43.956	12	82.09%
Tumor9	10.11	27.57%	2.552	9.08	28.03%	86.382	20	33.03%
Welsh	1.83	9.00%	.163	1.83	86.06%	11.488	8	90.24%

Table 10: Performance of the minimum gene subset selection on all datasets

The classification accuracy and the execution time of the greedy and set covering techniques were compared with Masked Sequential Backward selection technique for a gene subset comprising of 2-20 genes. The results are tabulated in Table 9. 50 repetitions of 4-fold cross validation was performed and the average values were computed. It can be found from the Table that minimum gene subset shows good performance for most of the gene set. For instance, on the Leukemia dataset, an almost maximum accuracy (87.00% vs 87.22%) is reached by the greedy selection using as few as 4.15 genes on average, whereas the maximum accuracy with a fixed subset is obtained by considering 20 genes. Independently of the dataset, the greedy minimum subset size is always larger than the set covering size. The greedy approach selects the gene maximizing the number of covered samples at each iteration. The set covering approach, instead, exploits a global optimization procedure to select the minimum number of genes that cover the samples. Hence, the greedy approach may need a larger number of genes to reach the best coverage of the training samples. This larger gene set provides a higher accuracy on most datasets, because it yields a more general model which may be less prone to over fitting. For instance, on the Leukemia dataset the average accuracy is 85.89% for the set covering approach and 87.00% for the greedy approach. The greedy algorithm is also characterized by a lower execution time with respect to the set covering algorithm. For example, considering the Brain2 dataset, the set covering completed in 35 seconds, whereas the greedy took less than 2 seconds. Since the greedy technique reaches higher classification accuracy with lower execution time, we have selected it as the method of choice both for the Masked sequential backward selection approach and for all the other experiments.

3.5. Classifier Bias

The effect of the peculiarities of different classification techniques on the gene set selected by sequential backward selection has been analyzed by comparing classification experiments performed with three different classifiers.

#	SB	IG	TR	SM	MM	GI	SV
2	84.39	82.89*	80.78*	82.08*	82.14*	83.39	**84.83**
4	**90.81**	88.67*	85.42*	88.83*	86.17*	89.81*	90.47
6	90.75	90.89	89.31*	91.11	87.03*	**91.19**	90.83
8	91.64	92.03	91.25	91.56	88.47*	**92.33**	91.64
10	92.08	92.78*	91.56	92.36	89.22*	**92.83***	92.25
12	92.75	92.86	92.03	92.81	90.42*	93.00	**93.19**
14	93.28	92.78	92.44*	92.47*	90.53*	92.89	**93.42**
16	93.94	92.58*	92.53*	92.31*	90.58*	92.44*	**94.03**
18	**94.67**	92.22*	92.33*	92.33*	90.44*	92.08*	94.19
20	**94.69**	92.47*	92.44*	92.31*	90.61*	92.31*	94.50
avg	91.90	91.02*	90.01*	90.82*	88.56*	91.23*	**91.94**
max	**94.69**	92.86	92.53	92.81	90.61	93.00	94.50
dev	2.85	2.97	3.72	3.11	2.63	2.77	2.72

Table 11: Accuracy obtained using KNN classifier on Leukemia dataset.

It is well known that different classification methods may classify the same dataset differently and produce a different classification performance. To know about the effect of different classification techniques on the gene set selected by Masked Sequential backward selection, the classification accuracy was compared with different classifiers. Three classifiers have been chosen as representatives of different classification techniques: (a) for decision trees, the J48 classifier of Weka [10], (b) the Support Vector Machine implemented in LibSVM [11], and, (c) for the K-Nearest Neighbors approach, the IBk implementation in Weka [10] with K=3. For LibSVM the provided script for automatically tuning the training phase (from data scaling to parameter selection) has been exploited, while for the other approaches the default parameter values have been set. The experiments have been performed on the Leukemia dataset for different numbers of selected features.

#	SB	IG	TR	SM	MM	GI	SV
2	84.42	84.19	82.44*	82.14*	82.47*	84.64	84.61
4	90.89	88.42*	86.33*	88.75*	86.81*	89.14*	90.36
6	91.31	90.08*	89.11*	90.06*	86.72*	90.39	90.56
8	91.47	90.92	90.67	90.56*	87.69*	91.14	91.11
10	92.28	90.75*	90.81*	91.58	88.92*	90.72*	91.94
12	92.89	91.17*	91.31*	91.39*	89.94*	91.22*	92.69
14	93.22	91.08*	91.58*	91.14*	90.92*	91.44*	93.25
16	93.39	91.64*	91.67*	91.97*	91.19*	91.69*	94.03
18	93.97	92.00*	91.47*	92.08*	91.33*	92.06*	94.75*
20	93.83	92.44*	91.67*	91.89*	91.39*	92.33*	94.72*
avg	91.77	90.27*	89.71*	90.16*	88.74*	90.48*	91.80
max	93.97	92.44	91.67	92.08	91.39	92.33	94.75
dev	2.66	2.28	2.89	2.85	2.72	2.13	2.84

Table 12: Accuracy obtained using SVM classifier on Leukemia dataset.

Table 11 and Table 12 show the results for the KNN and SVM classifiers respectively, while the results for the decision tree are reported in Table 3. It is clear that Masked Sequential Backward Selection and the Sum of Variance (SV) methods gives the best performance with similarity. Also, KNN and SVM show higher accuracy than that of the decision tree classifier. The first two classifiers build more robust models, which may make up for the selection of less interesting features by weighting them less in the model. Thus, decision trees allows better highlighting the effectiveness of different feature selection methods, because the quality of the selected feature set has a stronger impact on the accuracy obtained by the classifier. For this reason, we chose the decision tree to evaluate the quality of our feature selection method in the previous sections.

3.6. Computational Cost

The time required to classify a predefined number of genes was computed to be its computational cost. This feature was analyzed for sequential backward selection and other feature selection methods. We also analyzed the computational cost of our approach. We compared the time required by each approach to extract a high number of features (i.e.,1000 features) from the considered datasets. The Masked Sequential Backward Selection algorithm proved to be as efficient as the competing feature selection methods. In particular, on a Pentium 4 at 3.2 GHz with 2 GByte of RAM, the time required to extract the top 1000 genes on any complete dataset is in the order of few seconds (e.g., less than 1 second on the Alon dataset, 3 seconds on the Brain2 dataset) and very similar to the time required by the other methods.

4. Discussion

We analyzed the biological information presented in literature for the genes selected by the Masked Sequential Backward selection technique.

Rank	Gene ID	Gene Name	References
1	Z50753	GUCA2B	[12,13]
2	H06524	GSN	[14,13]
3	J02854	MYL9	[15,16,17,18,13]
4	K03474	AMH	[19]
5	L07032	PRKCQ	[20]
6	M63391	DES	[15,16,17,18,21,13]
7	M36634	VIP	[22,18,13]
8	R87126	MYH9	[23,18,13]
9	M76378	CSRP1	[15,16,17,18,13]
10	H43887	CFD	[15,16,17,13]
11	M22382	HSPD1	[15,16,17,13]
12	X63629	CDH3	[21,13]
13	H40095	MIFSLC2A11	[21,13]
14	X74295	ITGA7	[17]
15	T71025	MT1G	[13]
16	H77597	MT1G	[24]
17	J05032	DARS	[18]
18	X86693	SPARCL1	[17,18,13]
19	M26697	NPM1	[21,13]
20	H08393	OVGP1WDR77	[25]

Table 13: Top 20 genes on the Alon dataset (colon cancer) and related references.

Table 13 shows the first twenty genes selected by our algorithm on the entire Alon dataset, related to colon cancer and commonly used for biological validation [11, 26]. Most of the previous studies have validated the majority of the genes selected by masked sequential backward selection technique. Specifically the gene Z50753 is related to uroguanylin precursor [12], which may interfere with renewal and removal of epithelial cells. This will result in formation of polyps and ultimately malignant tumours of the colon and rectum [12]. Similarly the down regulation of the gene H06524 along with PRKCB1 leads to activation of Protein Kinase C (PKC) involved in phospholipid signaling and inhibit cell proliferation and tumorigenicity [14].

5. Conclusion

The proposed method Masked Sequential Backward selection technique on microarray data allows defining and ranking minimum set of genes with complete coverage of the training samples. The method has been compared with other feature selection techniques and it is found that the proposed method gives a better accuracy. Hence, the Masked Sequential Backward selection approach may provide a useful tool both to identify relevant genes for tumor diseases and to improve the classification accuracy of a classifier.

6. References

1. A. Statnikov, C.F. Aliferis, I. Tsamardinos, D. Hardin, and S. Levy.A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis.Bioinformatics, 21(5):631, 2005.
2. J.B. Welsh, L.M. Sapinoso, A.I. Su, S.G. Kern, J. Wang-Rodriguez, C.A. Moskaluk, H.F. Frierson, and G.M. Hampton. Analysis of gene expression identifies candidate markers and pharmacological targets in prostate cancer, 2001.
3. U. Alon, N. Barkai, DA Notterman, K. Gish, S. Ybarra, D. Mack, and AJ Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays.Proceedings of the National Academy of Sciences, 96(12):6745, 1999.
4. Y. Saeys, I. Inza, and P. Larranaga.A review of feature selection techniques in bioinformatics.Bioinformatics, 23(19):2507, 2007.
5. J. Hua, W.D. Tembe, and E.R. Dougherty.Performance of feature selection methods in the classification of high-dimension data.Pattern Recognition, 42(3):409–424, 2009.
6. S. Dudoit, J. Fridlyand, and T.P. Speed.Comparison of discrimination methods for the classification of tumors using gene expression data.Journal of the American Statistical Association, 97(457):77–88, 2002.
7. T. Li, C. Zhang, and M. Ogihara. A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression.Bioinformatics, 20(15):2429, 2004.
8. X. Zhou and D.P. Tuck. MSVM-RFE: extensions of SVM-RFE for multiclass gene selection on DNA microarray data. Bioinformatics, 23(9):1106, 2007.
9. I.H. Witten and E. Frank. Data Mining: Practical machine learning tools and techniques, 2005.
10. I.H. Witten and E. Frank. Data Mining: Practical machine learning tools and techniques, 2005. http://www.cs.waikato.ac.nz/ml/weka/
11. C.C. Chang and C.J. Lin. Training v-support vector classifiers: theory and algorithms, 2001. http://www.csie.ntu.edu.tw/~cjlin/libsvm
12. K. Shailubhai, H.H. Yu, K. Karunanandaa, J.Y. Wang, S.L. Eber, Y. Wang, N.S. Joo, H.D. Kim, B.W. Miedema, and S.Z. Abbas. Uroguanylin treatment suppresses polyp formation in the Apc Min/+ mouse and induces apoptosis in human colon adenocarcinoma cells via cyclic GMP. Cancer research, 60(18):5151–5157, 2000.

13. J.J. Chen, C.A. Tsai, S.L. Tzeng, and C.H. Chen.Gene selection with multiple ordering criteria.BMC bioinformatics, 8(1):74, 2007.
14. F. Bertucci, S. Salas, S. Eysteries, V. Nasserj, P. Finetti, C. Ginestier, E. Charafe-Jauffret, B. Loriod, L. Bachelart, and J. Montfort. Gene expression profiling of colon cancer by DNA microarrays and correlation with histoclinical parameters. Oncogene, 23(7):1377–1391, 2004.
15. Y.L. Yap, X.W. Zhang, M.T. Ling, X.H. Wang, Y.C. Wong, and A. Danchin. Classification between normal and tumor tissues based on the pair-wise gene expression ratio. BMC cancer, 4(1):72, 2004.
16. L. Wessels, M. Reinders, T. van Welsem, and P. Nederlof.Representation and classification for high-throughput data.In Proceedings of SPIE, volume 4626, page 226, 2002.
17. H. Kishino and P.J. Waddell.Correspondence analysis of genes and tissue types and finding genetic links from microarray data.GENOME INFORMATICS SERIES, pages 83–95, 2000.
18. M. Xiong, X. Fang, and J. Zhao.Biomarker identification by feature wrappers, 2001.
19. S. Wang, H. Chen, and S. Li. Gene Selection Using Neighborhood Rough Set from Gene Expression Profiles. In Proceedings of the 2007 International Conference on Computational Intelligence and Security, pages 959–963. IEEE Computer Society Washington, DC, USA, 2007.
20. W.H. Au, K.C.C. Chan, A.K.C. Wong, and Y. Wang. Attribute clustering for grouping, selection, and classification of gene expression data. IEEE/ACM transactions on computational biology and bioinformatics, pages 83–101, 2005.
21. S.M. Alladi, P. ShindeSantosh, V. Ravi, and U.S. Murthy.Colon cancjer prediction with genetic profiles using intelligent techniques.Bioinformation, 3(3):130, 2008.
22. Q. Tao, J. Ren, and J. Li. Vasoactive Intestinal Peptide Inhibits Adhesion Molecule Expression in Activated Human Colon Serosal Fibroblasts by Preventing NF-κB Activation.Journal of Surgical Research, 140(1):84–89, 2007.
23. W. Jiang, X. Li, S. Rao, L. Wang, L. Du, C. Li, C. Wu, H. Wang, Y. Wang, and B. Yang. Constructing disease-specific gene networks using pair-wise relevance metric: application to colon cancer identifies interleukin 8, desmin and enolase 1 as the central elements. BMC Systems Biology, 2(1):72, 2008.
24. C.P. Giacomini, S.Y. Leung, X. Chen, S.T. Yuen, Y.H. Kim, E. Bair, and J.R. Pollack. A gene expression signature of genetic instability in colon cancer, 2005.
25. G. Karakiulakis, C. Papanikolaou, SM Jankovic, A. Aletras, E. Papakonstantinou, E. Vretou, and V. Mirtsou-Fidani. Increased type IV collagen-degrading activity in metastases originating from primary tumors of the human colon.Invasion and metastasis, 17(3):158, 1997.
26. C. Furlanello, M. Serafini, S. Merler, and G. Jurman.Entropy-based gene ranking without selection bias for the predictive classification of microarray data.BMC bioinformatics, 4(1):54, 2003.