

Null space based feature selection method for gene expression data

Alok Sharma · Seiya Imoto · Satoru Miyano ·
Vandana Sharma

Received: 9 May 2011 / Accepted: 14 November 2011
© Springer-Verlag 2011

Abstract Feature selection is quite an important process in gene expression data analysis. Feature selection methods discard unimportant genes from several thousands of genes for finding important genes or pathways for the target biological phenomenon like cancer. The obtained gene subset is used for statistical analysis for prediction such as survival as well as functional analysis for understanding biological characteristics. In this paper we propose a null space based feature selection method for gene expression data in terms of supervised classification. The proposed method discards the redundant genes by applying the information of null space of scatter matrices. We derive the method theoretically and demonstrate its effectiveness on several DNA gene expression datasets. The method is easy to implement and computationally efficient.

Keywords Feature selection · Null space · DNA microarray gene expression data · Classification accuracy · Biological significance

1 Introduction

The feature selection methods have drawn widespread attention in the field of gene expression data analysis [3, 7, 9, 11–13, 17–19, 22–24, 26]. One of its important applications is in the human cancer classification. It provides the basis to identify crucial genes related to human cancers. The DNA microarray gene expression data are widely used for human cancer classification problem. The microarray gene expression data consist of large number of genes (dimensions) compared to the number of samples or feature vectors. The high dimensionality of the feature vectors degrades the generalization performance of the classifier and increases its computational complexity. This problem is popularly known as the small sample size (SSS) problem in the literature [10]. The feature selection method can be used here to retain only a few useful features and discard others, thereby, reducing the complexity in addition to finding important genes. Different approaches used in feature selection can be broadly grouped into two categories: filter approach and wrapper approach.¹ The filter approach is classifier independent whereas the wrapper approach is classifier dependent. Wrapper methods try to optimize the performance of a specific classifier and select the features yielding the best generalization performance. On the generalization, recently, some new techniques [28, 29, 31] based on maximizing the uncertainty or combining multiple reducts of rough sets have been proposed to improve the generalization of the classifier. In addition, some feature selection methods with genetic algorithm have been proposed in the literatures [5, 27].

A1 A. Sharma (✉) · S. Imoto · S. Miyano
A2 Laboratory of DNA Information Analysis, Human Genome
A3 Center, Institute of Medical Science, University of Tokyo,
A4 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan
A5 e-mail: aloks@ims.u-tokyo.ac.jp; sharma_al@usp.ac.fj

A6 A. Sharma
A7 School of Engineering and Physics,
A8 University of the South Pacific, Suva, Fiji

A9 V. Sharma
A10 CWM Hospital, Suva, Fiji

¹ The finer categorization of feature selection methods will include filter approach, wrapper approach and embedded approach [20].

56 In this paper we propose a feature selection method
 57 based on null linear discriminant analysis (LDA) technique
 58 [4]. The null LDA method is a feature extraction method,
 59 however, we have extended the notion to the feature
 60 selection domain. The proposed method falls under the
 61 wrapper approach category and is computationally effi-
 62 cient, simple and easy to implement. In this research, it has
 63 been applied on a number of DNA microarray gene
 64 expression data to show its classification effectiveness. In
 65 addition, the biological significance of the selected genes
 66 from the proposed method is also presented.

67 **2 Proposed gene selection method**

68 The null LDA technique [4] finds the orientation or
 69 transformation matrix \mathbf{W} in two stages. In the first stage,
 70 data is projected on the null space of within-class scatter
 71 matrix \mathbf{S}_W ; i.e., $\mathbf{S}_W \mathbf{W} = 0$. Then in the second stage it finds
 72 \mathbf{W} that maximizes the transformed between-class scatter
 73 matrix $|\mathbf{W}^T \mathbf{S}_B \mathbf{W}|$. The second stage is commonly imple-
 74 mented through the principal component analysis (PCA)
 75 method. Thus, in the null space method \mathbf{W} is found as

$$\mathbf{W} = \arg \max_{|\mathbf{W}^T \mathbf{S}_W \mathbf{W}|=0} |\mathbf{W}^T \mathbf{S}_B \mathbf{W}|$$

77 The orientation \mathbf{W} projects the vectors on the reduced
 78 feature space. The training vectors of a given class get
 79 merged into a single vector in this reduced feature space;
 80 i.e., class-conditional variances of the features in the
 81 reduced feature space are zero. The orientation $\mathbf{W} =$
 82 $[\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_h]$ has h column vectors, where $1 \leq h \leq c-1$
 83 and c is the number of class or state of the nature. When the
 84 dimensionality d of the original feature space is very large
 85 in comparison to the number of training vectors n , the
 86 evaluation of null space becomes nearly impossible as the
 87 eigenvalue decomposition of such a large $d \times d$ matrix
 88 leads to serious computational problems. However, there
 89 are methods available for efficiently computing the
 90 orientation \mathbf{W} [15, 30].

91 In order to introduce our gene selection method, let us
 92 first consider the two-class case. Let \mathcal{X} be a set of n training
 93 vectors in a d -dimensional feature space. The set \mathcal{X} can be
 94 subdivided into two subsets $\mathcal{X} = \{\mathbf{x}_1^1, \mathbf{x}_2^1, \dots, \mathbf{x}_{n_1}^1\}$ and $\mathcal{X} =$
 95 $\{\mathbf{x}_1^2, \mathbf{x}_2^2, \dots, \mathbf{x}_{n_2}^2\}$, where subset \mathcal{X}_i belongs to i -th class and
 96 consists of n_i number of samples such that $n = n_1 + n_2$;
 97 and the training vector \mathbf{x}_j^i represents j -th sample in i -th
 98 class. Let μ_i be the centroid of \mathcal{X}_i and μ be the centroid of \mathcal{X} .

99 Since the training vectors of a class get merged into a
 100 single vector in the reduced feature space, we can write
 101 training vectors of class-1 as:

$$y = \mathbf{w}^T \mathbf{x}_j^1 \quad \text{for } j = 1 \dots n_1, \tag{1}$$

Where $\mathbf{w} \in \mathbb{R}^d$ is any column vector of \mathbf{W} . Using Eq. 1,
 we get

$$\mathbf{w}^T (\mathbf{x}_1^1 - \mathbf{x}_2^1) = 0 \tag{2}$$

If we retain \mathbf{x}_1^1 in Eq. 2 and subtract it by the remaining
 samples of the same class then we will get in total $n_1 - 1$
 homogeneous equations; i.e., $\mathbf{w}^T (\mathbf{x}_1^1 - \mathbf{x}_j^1) = 0$ for $j = 2$
 $\dots n_1$. It can be seen from the Appendix section that the
 projection of any sample of a given class onto the null
 space of \mathbf{S}_W is independent of sample selection. Thus, we
 retain the first sample of a class without affecting the
 statistical stability. Now taking the average of these $n_1 - 1$
 equations, we get

$$\frac{1}{n} \mathbf{w}^T \left(n_1 \mathbf{x}_1^1 - \sum_{j=1}^{n_1} \mathbf{x}_j^1 \right) = 0$$

$$\text{or } \mathbf{w}^T \left[\frac{n_1}{n} \mathbf{x}_1^1 - \frac{n_1}{n} \mu_1 \right] = 0 \tag{3}$$

In a similar way, equation for class-2 can be written as

$$\mathbf{w}^T \left[\frac{n_2}{n} \mathbf{x}_1^2 - \frac{n_2}{n} \mu_2 \right] = 0 \tag{4}$$

Since mean of all the training vector can be given as
 $\mu = (n_1 \mu_1 + n_2 \mu_2) / n$, the summation of Eqs. 3 and 4 will
 give

$$\mathbf{w}^T (\bar{\mathbf{x}} - \mu) = 0, \tag{5}$$

where $\bar{\mathbf{x}} = (n_1 \mathbf{x}_1^1 + n_2 \mathbf{x}_1^2) / n$. Eq. 5 can be written for all
 the d features in the summation form as

$$\sum_{j=1}^d w_j (\bar{x}_j - \mu_j) = 0 \tag{6}$$

where w_j , \bar{x}_j and μ_j are the features of \mathbf{w} , $\bar{\mathbf{x}}$ and μ
 respectively. It can be observed from Eq. 6 that if the j -th
 component is very small; i.e., $|w_j (\bar{x}_j - \mu_j)| \approx 0$, it will not
 contribute much in making the overall sum equal to zero
 and thus the j -th component or j -th feature can be discarded
 without sacrificing much information. If $|w_j (\bar{x}_j - \mu_j)|$ is
 arranged in descending order such that

$$|w_1 (\bar{x}_1 - \mu_1)| \geq |w_2 (\bar{x}_2 - \mu_2)| \geq \dots \geq |w_d (\bar{x}_d - \mu_d)|$$

then by discarding the bottom features (for which
 $|w_j (\bar{x}_j - \mu_j)| \approx 0$) will not affect much the qualitative
 importance of the retained top r features.

The two-class case can be easily extended to the multi-
 class case. If there are c classes and the number of training
 vectors per class is n_i where $i = 1, 2, \dots, c$ then $\bar{\mathbf{x}}$ in Eq. 5
 can be computed as

Author Proof

$$\bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^c n_i \mathbf{x}_i^i \quad \text{where } n = \sum_{i=1}^c n_i \quad (7)$$

141 and μ is the mean of all the training vectors. We derived the
 142 method by using the first training vector from each class.
 143 However, one can take any training vector from each of the
 144 classes.

145 The selected number of features should be in the range
 146 $n - c < r < d$, as $r \leq n - c$ will make the within-class
 147 scatter matrix non-singular and null LDA method cannot be
 148 applied on that case. There are several ways of computing the
 149 value of r . One way is to find the argument of median of
 150 $|w_j(\mathbf{x}_j - \mu_j)|$ for $j = 1 \dots d$; i.e., $r_1 = \arg \text{median}$
 151 $_{j=1 \dots d} [|w_j(\mathbf{x}_j - \mu_j)|]$, this will discard approximately 50%
 152 features. The reduction can be continued further as
 153 $r_2 = \arg \text{median}_{j=1 \dots r_1} [|w_j(\mathbf{x}_j - \mu_j)|]$. This procedure will
 154 discard approximately 75% of features. Alternatively, cross-
 155 validation procedure (illustrated later in this section) can be
 156 applied to find the optimum r value from the training data set.
 157 The method is summarized in Table 1.

158 The proposed gene selection method does not evaluate
 159 the performance of every single gene. It takes all the genes
 160 and evaluates the best genes. Therefore, the genes can be
 161 selected with very low computational complexity, $O(dn^2)$.
 162 The effectiveness of the method is discussed in the
 163 experimentation section.

164 If the optimum value of r is required then k -fold cross
 165 validation procedure can be used [21]:

- 166 *Step 1* Partition training data randomly into k roughly
 167 equal segments.
- 168 *Step 2* Hold out one segment as validation data and the
 169 remaining $k - 1$ segments as learning data from the
 170 training data.
- 171 *Step 3* Select r (where $n - c < r < d$) and apply the
 172 proposed method (shown in Table 1) using learning data
 173 to find top r features.
- 174 *Step 4* Use validation data to compute classification
 175 accuracy for a range of values of r . Store the obtained
 176 classification accuracies.
- 177 *Step 5* Repeat steps 1–4 N times.
- 178 *Step 6* Evaluate average classification accuracy over
 179 N repetitions.
- 180 *Step 7* Plot a curve of average classification accuracy as
 181 a function of r .

Table 1 Gene selection method

Step 1. Compute orientation or transformation vector $\mathbf{w} \in \mathbf{W}$ using null LDA method
Step 2. Compute $\bar{\mathbf{x}}$ using Eq. 7 and mean of training vectors μ
Step 3. Arrange $ w_j(\mathbf{x}_j - \mu_j) $ in descending order for $j = 1 \dots d$
Step 4. Select the top r features

Step 8 The argument of maximum average classification accuracy will be the optimum r value. 182
183

3 Experimentation 184

Four DNA microarray gene expression datasets² are utilized in this work to show the effectiveness of the proposed method. The description of the datasets is given as follows: 185
186

Acute leukemia dataset [12, 13]: this dataset consists of DNA microarray gene expression data of human acute leukemia for cancer classification. Two types of acute leukemia data are provided for classification namely acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML). The dataset is subdivided into 38 training samples and 34 test samples. The training set consists of 38 bone marrow samples (27 ALL and 11 AML) over 7,129 probes. The testing set consists of 34 samples with 20 ALL and 14 AML, prepared under different experimental conditions. All the samples have 7,129 dimensions and all are numeric. 187
188
189
190
191
192
193
194
195
196
197
198
199

SRBCT dataset [16]: the small round blue-cell tumor dataset consists of 83 samples with each having 2,308 genes. This is a four class classification problem. The tumors are *Burkitt lymphoma* (BL), *the Ewing family of tumors* (EWS), *neuroblastoma* (NB) and *rhabdomyosarcoma* (RMS). There are 63 samples for training and 20 samples for testing. The training set consists of 8, 23, 12 and 20 samples of BL, EWS, NB and RMS respectively. The test set consists of 3, 6, 6 and 5 samples of BL, EWS, NB and RMS respectively. 200
201
202
203
204
205
206
207
208
209

MLL leukemia dataset [2]: this dataset has three classes namely ALL, MLL and AML. The training set contains 57 leukemia samples (20 ALL, 17 MLL and 20 AML) whereas the test set contains 15 samples (4 ALL, 3 MLL and 8 AML). The dimension of MLL dataset is 12,582. 210
211
212
213
214

Lung Dataset [14]: this dataset contains gene expression levels of malignant mesothelioma (MPM) and adenocarcinoma (ADCA) of the lung. There are 181 tissue samples (31 MPM and 150 ADCA). The training set contains 32 of them, 16 MPM and 16 ADCA. The rest of 149 samples are used for testing. Each sample is described by the expression values of 12,533 genes. 215
216
217
218
219
220
221

In order to study the performance of the proposed feature selection method, we first evaluate the classification 222
223

² Most of the datasets are downloaded from the Kent Ridge Biomedical Dataset (KRBD) (<http://datam.i2r.a-star.edu.sg/datasets/krbd/>). The datasets are transformed or reformatted and made available by KRBD repository and we have used them without any further preprocessing. Some datasets which are not available on KRBD repository are downloaded and directly used from respective authors' supplement link. The URL addresses for all the datasets are given in the Reference Section. 2FL01
2FL02
2FL03
2FL04
2FL05
2FL06
2FL07
2FL08

Table 2 Comparison between the classification accuracy obtained using the selected features (by the proposed method) and using original feature space

Dataset	Number of genes						
	100 (%)	400 (%)	500 (%)	750 (%)	1,000 (%)	2,000 (%)	Dim (%)
SRBCT	95.00	95.00	100.00	100.00	100.00	100.00	100.00 (dim = 2,308)
Acute leukemia	100.00	97.06	97.06	97.06	97.06	97.06	97.06 (dim = 7,129)
MLL leukemia	100.00	100.00	100.00	100.00	100.00	100.00	100.00 (dim = 12,582)
Lung cancer	97.32	97.99	97.99	97.99	97.99	97.99	97.99 (dim = 12,533)

224 accuracy on the original feature set using null LDA method
 225 and nearest neighbor classifier (with Euclidean distance
 226 measure). Then we select the features using the proposed
 227 method and apply null LDA method and nearest neighbor
 228 classifier to see if there is any degradation in the classifi-
 229 cation accuracy. The results are summarized in Table 2.

230 It can be observed from Table 2 that for most of the
 231 datasets there is no degradation in classification accuracy
 232 until the number of features is reduced to 500, which is a
 233 fair amount of reduction. However, for MLL leukemia there
 234 is no degradation at all in the performance even when the
 235 number of features is reduced to 100; i.e., after 99.2%
 236 feature reduction. On the other hand, for acute leukemia
 237 dataset the classification accuracy is actually improved to
 238 the perfect level (100%) when only 100 genes (1.4% of the
 239 original dimension) are selected. This shows that some
 240 unimportant genes are discarded which helps in improving
 241 the classification performance. In a similar way, the clas-
 242 sification accuracy for SRBCT dataset utilizing only 100
 243 genes is recorded at 95% (using 4.33% of the original
 244 dimension). Furthermore, on Lung Cancer dataset the
 245 classification accuracy of selected 100 genes is recorded at
 246 97.32% (using 0.79% of the original dimension) which is a
 247 fraction less when compared with the classification accu-
 248 racy of original dimension. It can be concluded from the
 249 table that a large amount of unimportant genes can be dis-
 250 carded without significant loss of discriminant information.

251 Table 3 shows the comparison between the classifica-
 252 tion accuracy with the other feature selection algorithms
 253 [25] on SRBCT and MLL leukemia datasets. It can be
 254 observed from Table 3 that the null space based feature
 255 selection method outperforms many other state-of-the-art
 256 feature selection methods. The proposed method achieves
 257 95% and 100% classification accuracies on SRBCT using
 258 100 genes and 500 genes, respectively. Some of the feature
 259 selection methods (like information gain + SVM random,
 260 towing rule + SVM random etc.) also achieve 95% clas-
 261 sification accuracy but using 150 genes. Similarly, the
 262 proposed method achieves 100% classification accuracy on
 263 MLL leukemia dataset using only 100 genes.

264 Table 4 shows the performance in terms of classification
 265 accuracy on acute leukemia dataset. The proposed method

266 achieves 100% classification accuracy using 100 genes and
 267 is better than all the other presented methods. The pre-
 268 diction strength + SVM achieves between 88.2 and 94.1%
 269 using 25–1,000 genes. RCBT, discretization + decision
 270 tree and rough sets achieve optimum classification accu-
 271 racy of 91.2% using 10–40, 1,038 and 9 genes, respec-
 272 tively. Though some of these methods use small number of
 273 genes, they are computationally intensive.

274 Table 5 shows the performance on Lung Cancer dataset.
 275 The null space based feature selection method exhibits
 276 97.3% classification accuracy. The classification accuracy
 277 on Lung Cancer dataset is 0.6% lower than RCBT and PCL
 278 methods. However, it can be made equal by increasing the
 279 number of selected genes.

280 In general, it can be concluded from the Tables 3, 4, 5
 281 that the null space based feature selection method exhibits
 282 promising results.

283 In order to assess the reliability of the null space based
 284 feature selection method, we conducted sensitivity testing
 285 [1, 6] on test data. The sensitivity of a class is defined by
 286 the number of true positives over the number of true posi-
 287 tives + the number of false negatives. Table 6 shows the
 288 sensitivity of the method on all the four datasets. In the
 289 table s_i (for $j = 1 \dots c$) denotes the sensitivity of class i ,
 290 $s = \min(s_i)$ is the minimum of sensitivity and $g =$
 291 $(\prod_{i=1}^c s_i)^{1/c}$ is the c -th root of the product of s_i , where
 292 $0 \leq s_i \leq 1$. The minimum sensitivity s can be considered as
 293 a complementary measure of classification accuracy whose
 294 value should be high for a good method [6]. The term g can
 295 be considered as an imbalance of classification accuracy
 296 among the classes [1]. It can be observed from the table
 297 that both the terms (s and g) for the null space based feature
 298 selection method on acute leukemia and MLL leukemia
 299 datasets achieve perfect results; for Lung Cancer dataset
 300 the method misclassified four test samples in class 2 and
 301 for SRBCT dataset the method misclassified one test
 302 sample in class 3. Overall the high values of these terms
 303 depict that the method's ability to identify cancer classes is
 304 very reliable.

305 Moreover, it would be interesting to see the biological
 306 significance of the selected features by the null based

Table 3 Comparison of feature selection methods on SRBCT and MLL leukemia datasets

Methods (feature selection + classification)	# Selected genes	SRBCT (classification accuracy) (%)	MLL leukemia (classification accuracy) (%)
Information gain + Naïve Bayes	150	68	66
Information gain + SVM random	150	95	100
Information gain + SVM exhaustive	150	91	100
Towing rule + Naïve Bayes	150	73	86
Towing rule + SVM random	150	95	100
Towing rule + SVM exhaustive	150	95	100
Sum minority + Naïve Bayes	150	68	26
Sum minority + SVM random	150	95	86
Sum minority + SVM exhaustive	150	91	80
Max minority + Naïve Bayes	150	77	34
Max minority + SVM random	150	91	86
Max minority + SVM exhaustive	150	91	80
Gini index + SVM Naïve Bayes	150	78	66
Gini index + SVM random	150	95	100
Gini index + SVM exhaustive	150	95	100
Sum of variances + SVM Naïve Bayes	150	63	54
Sum of variances + SVM random	150	91	100
Sum of variances + SVM exhaustive	150	95	100
t-statistics + Naïve Bayes	150	63	54
t-statistics + SVM random	150	91	100
t-statistics + SVM exhaustive	150	95	100
One-dimensional SVM + SVM Naïve Bayes	150	63	54
One-dimensional SVM + SVM random	150	91	100
One-dimensional SVM + SVM exhaustive	150	95	100
Proposed feature selection + null LDA with nearest centroid classifier	500	100.00	100
Proposed feature selection + null LDA with nearest centroid classifier	100	95	100

Table 4 Comparison of feature selection methods on acute leukemia dataset

Methods (feature selection + classification)	# Selected genes	Classification accuracy (%)
Prediction strength + SVM [11]	25–1,000	88.2–94.1
Neighborhood analysis + weighted voting [12, 13]	50	85.3
RCBT [7]	10–40	91.2
Discretization + decision trees [24]	1,038	91.2
Rough sets, GA + kNN [3]	9	91.2
Proposed feature selection + null LDA with nearest centroid classifier	100	100

Table 5 Comparison of feature selection methods on lung cancer dataset

Methods (feature selection + classification)	# Selected genes	Classification accuracy (%)
RCBT [7]	10–40	98
Discretization + decision trees [24]	5,365	93
Boosting [17]	Unknown	81
Bagging [17]	Unknown	88
PCL [17]	Unknown	98
Proposed feature selection + null LDA with nearest centroid classifier	100	97.3

307 feature selection algorithm. We use acute leukemia data as
 308 a prototype to show the biological significance using
 309 Ingenuity Pathway Analysis.³ The selected 100 features

from the proposed algorithm are used for this purpose. The
 top five high level biological functions obtained are shown
 in Fig. 1. In the figure, the y-axis denotes the negative of
 logarithm of p-values and x-axis denotes the high level
 functions.

310
 311
 312
 313
 314

3FL01 ³ IPA, <http://www.ingenuity.com>.

Table 6 Sensitivity testing on all the four datasets

Datasets	s_1	s_2	s_3	s_4	s (in percentage)	g
Acute leukemia	1	1	–	–	100.0	1
Lung cancer	1	130/134	–	–	97.0	0.9850
MLL leukemia	1	1	1	–	100.0	1
SRBCT	1	1	5/6	1	83.3	0.9554

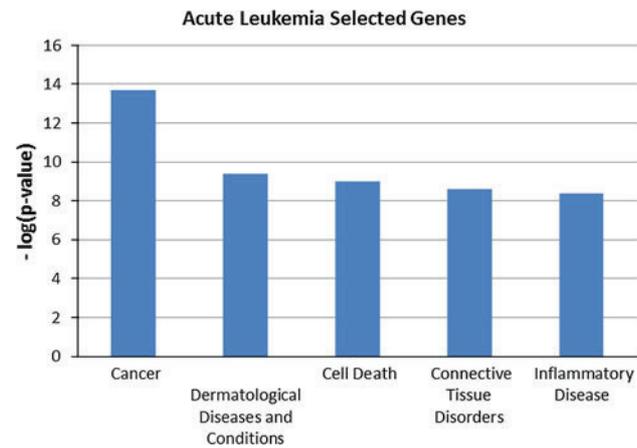


Fig. 1 Top five high level biological functions on selected 100 features of acute leukemia by the null space based feature selection method

Table 7 Cancer functions

Functions	p-value	# Selected genes
Cancer	2.20E-14	48
Leukemia	2.70E-03	7
Lymphocytic leukemia	3.46E-03	5
Acute myeloid leukemia	7.79E-03	4
Acute lymphocytic leukemia	1.21E-02	3
Type M6 acute myeloid leukemia	1.55E-02	1
Acute monocytic leukemia (M5)	2.06E-02	1

315 Since the cancer function is of paramount interest, we
 316 investigated them further. There are 66 cancer functions
 317 obtained from the experiment. The leukemia is selected
 318 from these 66 cancer functions and shown in Table 7. In
 319 the table, the p-values and the number of selected genes are
 320 depicted corresponding to the selected functions. The
 321 selected genes by the proposed method provide significant
 322 p-values above the threshold (as specified in IPA). This
 323 shows that the features selected by the proposed method
 324 contain useful information for discriminatory purpose and
 325 have biological significance.

326 In order to check the robustness of the proposed method,
 327 we carried out sensitivity analysis. First we selected top

Table 8 Sensitivity analysis for null space based feature selection method on gene expression datasets at different noise levels

Added noise (%)	SRBCT (%)	MLL leukemia (%)	Acute leukemia (%)	Lung cancer (%)
1	89.6 (65)	95.2 (65)	94.1 (82)	97.5 (79)
2	88.7 (65)	91.2 (40)	91.3 (60)	97.1 (62)
5	89.9 (66)	59.5 (30)	62.4 (49)	89.3 (51)

The average classification accuracy is shown in percentage and the average number of common genes is shown in parenthesis

100 genes using the proposed method on a given dataset. 328
 After this we contaminated the dataset by adding Gaussian 329
 noise; then applied the method again to find the top 100 330
 genes. The generated noise levels are 1, 2 and 5% of the 331
 standard deviation of the original expression values. The 332
 number of genes which are common after contamination 333
 and before contamination is noted. In addition, the classi- 334
 fication accuracy is also noted. This contamination, selec- 335
 tion of genes and computation of classification accuracy is 336
 repeated 50 times. The average number of genes and 337
 average classification accuracy (in percentage) over 50 338
 iterations is depicted in Table 8. It can be observed from 339
 the table that the proposed method can achieve high clas- 340
 sification accuracy in an adverse environmental or noisy 341
 condition. Also, the method is able to capture the majority 342
 of original genes in the noisy condition. 343

4 Conclusion 344

In this paper we proposed the null space based feature 345
 selection method. The proposed method effectively selects 346
 important genes which have been demonstrated on several 347
 DNA microarray gene expression data. Comparisons with 348
 several other feature selection methods have shown that the 349
 proposed method has better classification accuracy. The 350
 implementation of the method is also quite simple and the 351
 computation is fast. Finally, the selected genes by the 352
 proposed method have biological significance which is 353
 demonstrated by performing functional analysis and will 354
 therefore contribute positively towards detection of sig- 355
 nificant biological phenomenon. 356

Acknowledgments We thank the Reviewers and the Editor for their 357
 constructive comments which appreciably improved the presentation 358
 quality of the paper. 359

Appendix 360

Theorem 1 Let the column vectors of orthogonal matrix 361
 \mathbf{W} span the null space of within-class scatter matrix \mathbf{S}_W 362

363 and $\mathbf{w} \in \mathbb{R}^d$ be any column vector of \mathbf{W} . Let the j -th
 364 sample in i -th class be denoted by $\mathbf{x}_j^i \in \mathbb{R}^d$. Then the pro-
 365 jection of sample \mathbf{x}_j^i onto the null space of \mathbf{S}_W is inde-
 366 pendent of the sample selection in class.

367 *Proof* Since $\mathbf{w} \in \mathbb{R}^d$ is in the null space of \mathbf{S}_W , by def-
 368 inition $\mathbf{S}_W \mathbf{w} = 0$ or $\mathbf{w}^T \mathbf{S}_W \mathbf{w} = 0$. The within-class scatter
 369 matrix \mathbf{S}_W is a sum of scatter matrices $\mathbf{S}_W = \sum_{i=1}^c \mathbf{S}_i$,
 370 where c denotes the number of classes and scatter matrix \mathbf{S}_i
 371 can be represented by [8]:

$$\mathbf{S}_i = \sum_{j=1}^{n_i} (\mathbf{x}_j^i - \boldsymbol{\mu}_i) (\mathbf{x}_j^i - \boldsymbol{\mu}_i)^T \quad (\text{A1})$$

373 where $\boldsymbol{\mu}_i$ denotes the mean of class i and n_i denotes the
 374 number of samples in class i .

375 Since $\mathbf{w}^T \mathbf{S}_W \mathbf{w} = 0$ (or $\mathbf{w}^T \sum_{i=1}^c \mathbf{S}_i \mathbf{w} = 0$) and \mathbf{S}_i is
 376 positive semi-definite matrix, we can represent $\mathbf{w}^T \mathbf{S}_i \mathbf{w} = 0$.
 377 From Eq. A1, we can say

$$\begin{aligned} & \mathbf{w}^T \sum_{j=1}^{n_i} (\mathbf{x}_j^i - \boldsymbol{\mu}_i) (\mathbf{x}_j^i - \boldsymbol{\mu}_i)^T \mathbf{w} = 0 \\ \text{or } & \sum_{j=1}^{n_i} \mathbf{w}^T \mathbf{x}_j^i \mathbf{x}_j^{iT} \mathbf{w} - \sum_{j=1}^{n_i} \mathbf{w}^T \boldsymbol{\mu}_i \boldsymbol{\mu}_i^T \mathbf{w} \\ \text{or } & \sum_{j=1}^{n_i} \left(\left\| \mathbf{w}^T \mathbf{x}_j^i \right\|^2 - \left\| \mathbf{w}^T \boldsymbol{\mu}_i \right\|^2 \right) = 0 \end{aligned} \quad (\text{A2})$$

379 where $\|\cdot\|$ is the Euclidean norm. Eq. A2 immediately
 380 leads to $\mathbf{w}^T \mathbf{x}_j^i = \mathbf{w}^T \boldsymbol{\mu}_i$; i.e., projection of sample \mathbf{x}_j^i onto the
 381 null space of \mathbf{S}_W is independent of j (or in other words
 382 independent of sample selection). This concludes the proof
 383 of the Theorem.
 384

385 **References**

386 1. Arif M, Akram MU, Minhas FAA (2010) Pruned fuzzy k-nearest
 387 neighbor classifier for beat classification. J Biomed Sci Eng
 388 3:380–389
 389 2. Armstrong SA, Staunton JE, Silverman LB, Pieters R, den Boer
 390 ML, Minden MD, Sallan SE, Lander ES, Golub TR, Korsmeyer
 391 SJ (2002) MLL translocations specify a distinct gene expression
 392 profile that distinguishes a unique leukemia. Nature Genetics
 393 30:41–47 (Data Source 1: [http://datam.i2r.a-star.edu.sg/datasets/](http://datam.i2r.a-star.edu.sg/datasets/krbd/)
 394 [krbd/](http://www.broad.mit.edu/cgi-bin/cancer/publications/pub_paper.cgi?mode=view&paper_id=63)) (Data Source 2: [http://www.broad.mit.edu/cgi-bin/cancer/](http://www.broad.mit.edu/cgi-bin/cancer/publications/pub_paper.cgi?mode=view&paper_id=63)
 395 [publications/pub_paper.cgi?mode=view&paper_id=63](http://www.broad.mit.edu/cgi-bin/cancer/publications/pub_paper.cgi?mode=view&paper_id=63))
 396 3. Banerjee M, Mitra S, Banka H (2007) Evolutionary-rough feature
 397 selection in gene expression data. IEEE Trans Syst Man Cybern
 398 Part C Appl Rev 37:622–632
 399 4. Chen L-F, Liao H-YM, Ko M-T, Lin J-C, Yu G-J (2000) A new
 400 LDA-based face recognition system which can solve the small
 401 sample size problem. Pattern Recognit 33:1713–1726
 402 5. Boehm O, Haroon DR, Manevitz LM (2011) Classifying cog-
 403 nitive states of brain activity via one-class neural networks with
 404 feature selection by genetic algorithms. Int J Mach Learn Cybern
 405 2(3):125–134
 406 6. Caballero JCF, Martinez FJ, Hervas C, Gutierrez PA (2010)
 407 Sensitivity versus accuracy in multiclass problems using memetic

Pareto evolutionary neural networks. IEEE Trans Neural Netw
 21(5):750–770 408
 409
 7. Cong G, Tan K-L, Tung AKH, Xu X (2005) Mining top-k cover-
 ing rule groups for gene expression data. In: The ACM SIG-
 MOD International Conference on Management of Data, 411
 pp 670–681 412
 8. Duda RO, Hart PE (1973) Pattern classification and scene anal-
 ysis. Wiley, New York 414
 9. Dudoit S, Fridlyand J, Speed TP (2002) Comparison of dis-
 criminant methods for the classification of tumors using gene
 expression data. J Am Stat Assoc 97:77–87 416
 10. Fukunaga K (1990) Introduction to statistical pattern recognition.
 Academic Press Inc., Hartcourt Brace Jovanovich, Publishers, 417
 Boston 418
 11. 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):906–914 419
 12. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M,
 Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA et al
 (1999) Molecular classification of cancer: class discovery and class
 prediction by gene expression monitoring. Science 286(5439):
 531–537 420
 421
 13. 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 286:531–537 (Data Source: <http://datam.i2r.a-star.edu.sg/datasets/krbd/>)
 422
 423
 14. Gordon GJ, Jensen RV, Hsiao L-L, Gullans SR, Blumenstock JE,
 Ramaswamy S, Richards WG, Sugarbaker DJ, Bueno R (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 (Data Source 1: [http://](http://datam.i2r.a-star.edu.sg/datasets/krbd/)
 424
 425
 426
 427
 428
 429
 430
 431
 432
 433
 434
 435
 436
 437
 438
 439
 440
 441
 442
 443
 444
 445
 446
 447
 448
 449
 450
 451
 452
 453
 454
 455
 456
 457
 458
 459
 460
 461
 462
 463
 464
 465
 466
 467
 468
 469
 470
 471
 472

Author Proof

- 473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
24. Tan AC, Gilbert D (2003) Ensemble machine learning on gene expression data for cancer classification. *Appl Bioinforma* 2(3 Suppl):S75–S83
25. Tao L, Zhang C, Ogihara M (2004) A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. *Bioinformatics* 20(14): 2429–2437
26. Thomas J, Olson JM, Tapscott SJ, Zhao LP (2001) An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles. *Genome Res* 11:1227–1236
27. Tong DL, Mintram R (2010) Genetic Algorithm-Neural Network (GANN): a study of neural network activation functions and depth of genetic algorithm search applied to feature selection. *Int J Mach Learn Cybern* 1(1–4):75–87
28. Wang X-Z, Dong C-R (2009) Improving generalization of fuzzy if-then rules by maximizing fuzzy entropy. *IEEE Trans Fuzzy Syst* 17(3):556–567
29. Wang X-Z, Zhai J-H, Lu S-X (2008) Induction of multiple fuzzy decision trees based on rough set technique. *Inf Sci* 178(16): 3188–3202
30. Ye J (2005) Characterization of a family of algorithms for generalized discriminant analysis on under sampled problems. *J Mach Learn Res* 6:483–502
31. Zhao H-X, Xing H-J, Wang X-Z (2011) Two-stage dimensionality reduction approach based on 2DLDA and fuzzy rough sets technique. *Neurocomputing* 74:3722–3727
- 488
489
490
491
492
493
494
495
496
497
498
499
500
501

UNCORRECTED PROOF