SVM-RFE-ED: A Novel SVM-RFE based on Energy Distance for Gene Selection and Cancer Diagnosis

Seyyid Ahmed Medjahed¹,², Mohammed Ouali³

¹ Centre Universitaire Ahmed Zabana Relizane, Algeria
² Université des Sciences et de la Technologie d’Oran Mohamed Boudiaf, Algeria
³ College of Computers and Information Technology, Taif University, Kingdom of Saudi Arabia

seyyid.ahmed@univ-usto.dz, m.ouali@tu.edu.sa

Abstract. Microarray expression data has been a very active research field and an indispensable tool for cancer diagnosis. The microarray expression dataset contains thousands of genes and selecting a subset of informative genes is a primordial preprocessing step for improving the cancer classification. Support Vector Machine Recursive Feature Elimination (SVM-RFE) is one of the popular and effective gene selection approaches. However, SVM-RFE attempts to find the best possible combination for classification and does not take into account the ability of class separability for each gene. In this paper, a novel SVM-RFE based on energy distance (ED) and called SVM-RFE-ED is proposed to overcome the limitation of standard SVM-RFE. The aims of our study are to achieve a high classification accuracy rate and improve the classification model. The experimentation is conducted on five widely used datasets. Experimental results indicate that the proposed approach SVM-RFE-ED provides good results and achieve a high classification accuracy rate using a small number of genes.

Keywords. Cancer diagnosis, support vector machine, recursive feature elimination, gene selection, energy distance, classification.

1 Introduction

Feature selection has been a very active research field in many application [16, 14, 17]. Recently, DNA microarray technology has gained attention from biologists and scientists to improve the process of cancer diagnosis [8, 10]. DNA microarray datasets are composed of large number of genes expression and a few dozen of instances. This characteristic increases the risk of overfitting in the classification process and reduces significantly the quality of the classification model. In order to overcome this problem, it is very important to reduce the number of genes by selecting the informative subset of genes and eliminating the irrelevant and redundant genes. This preprocessing phase is called gene selection.

Gene selection or feature selection aims to select the smallest subset of genes without reduces the classification accuracy rate. It can be divided into three classes: the first one is the filter feature selection approach and it evaluates the candidate subset of genes independently of the classifier system. The second one is the wrapper approach and it uses the classifier system to compute the fitness of genes subset. The last one is the embedded feature selection approach and it incorporates the gene selection procedure in the classification system.

In this paper, we propose a novel SVM-RFE approach called SVM-RFE-ED that incorporates the energy distance to compute the class separability and to minimize the number of genes. This approach aims to select the smallest subset of
genes that provides a high classification accuracy rate. The performance assessment is demonstrated on five datasets used for cancer diagnosis *i.e.*, colon dataset, leukemia dataset, lung dataset, ovarian dataset and DLBC dataset.

Experimental results indicate that the proposed approach SVM-RFE-ED produces very satisfactory results and a high classification accuracy rate. The stability of the proposed approach was demonstrated.

The rest of paper is organized as follows: In Section 2, we present and detailed the proposed approach. In Section 3, the results are critically analyzed with the existing approaches. Finally, in section 4, the conclusion and some perspectives are given.

## 2 The Proposed Approach

**SVM-RFE-ED**

### 2.1 SVM-RFE Algorithm

SVM-RFE (Support Vector Machine - Recursive Feature Elimination), is an iterative algorithm that ranks the initial genes according to score function and eliminates the genes with the lowest scores. SVM-RFE is proposed by Guyon et al. [6], the basic idea is to train the algorithm by using SVM with some kernel function and recursively eliminates the genes using the smallest ranking score [9].

SVM [2] is one of the popular kernel-based approach used to classify the data. Mathematically, for some dataset $D = \{(x_1, y_1), ... , (x_N, y_N)\}$, $x \in \mathbb{R}^n$, $y \in \{-1, 1\}$, SVM attempts to find the optimal hyperplane that separates two classes by maximizing the margin (primal problem):

\[
\begin{align*}
\min_{w,b,\xi} & \quad \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{N} \xi_i, \\
\text{subject to} & \quad y_i(w \cdot x_i) + b \geq 1 - \xi_i, \\
& \quad \xi_i \geq 0, \\
& \quad i \in \{1, ..., N\},
\end{align*}
\]

(1)

where $\xi_i$ mesure the degree of misclassification of the data $x_i$ and $C$ is the regularization parameter which controls the trade-off between the percentage of misclassified and the size of the margin [2].

The dual problem of (1) is given as follows:

\[
\begin{align*}
\min_\alpha & \quad - \sum_{i=1}^{N} \alpha_i + \frac{1}{2} \sum_{i,j} y_i y_j \alpha_i \alpha_j k(x_i, x_j), \\
\text{subject to} & \quad \sum_{i=1}^{N} \alpha_i y_i = 0, \\
& \quad \forall i \in \{1, ..., N\}, \quad 0 \leq \alpha_i \leq C,
\end{align*}
\]

(2)

where $k(x_i, x_j)$ is the kernel function. Kernel functions allow a nonlinear transformation of data into a linear separation of examples in a new space called “feature space” which has many dimensions. Several kernel functions have been defined in the literature, the most used are described in Table 1, where $\alpha_i$ is the solution of the dual problem. The primal solution is given as follows:

\[
\begin{align*}
w^* &= \sum_{i=1}^{N} \alpha_i y_i x_i, \\
b^* &= -\frac{1}{2} \langle w^*, x_r + x_s \rangle.
\end{align*}
\]

### Table 1. The most used kernel functions

<table>
<thead>
<tr>
<th>Kernel name</th>
<th>Formulation</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear</td>
<td>$k(x, y) = x \cdot y$</td>
<td>$l$</td>
</tr>
<tr>
<td>Polynomial</td>
<td>$k(x, y) = (x \cdot y)^d$</td>
<td>$d$</td>
</tr>
<tr>
<td>Gaussian</td>
<td>$k(x, y) = \exp(-\frac{|x - y|^2}{\sigma^2})$</td>
<td>$\sigma$</td>
</tr>
<tr>
<td>Multilayer Perceptron</td>
<td>$k(x, y) = \tan(P_1 x^t y + P_2)$</td>
<td>$P_1, P_2$</td>
</tr>
<tr>
<td>Quadratic</td>
<td>$k(x, y) = (x \cdot y + 1)^2$</td>
<td>$l$</td>
</tr>
</tbody>
</table>

SVM-RFE uses the coefficient vector $w_i$ to generate a rank:

\[
w_i = \sum_{i=1}^{n} \alpha_i x_i y_i,
\]

$\alpha_i$ is the Lagrangian Multiplier, $x_i$ is the gene expression, $y_i$ is the class.

The general schema of SVM-RFE algorithm can be described as in Algorithm 1.

Unfortunately, the performance of SVM-RFE become unstable at some values of the gene filter out *i.e.* the number of gene eliminate in each iteration [11]. In addition, SVM-RFE find a combination for classification and does not take
The energy distance is a statistic distance between the distributions of random vectors. The origin of the name “energy” is taken from Newton gravitational potential energy and is based from the distance between two bodies [15]:

\[
\varepsilon(X, Y) = \frac{2}{nm^2} \sum_{i=1}^{n_1} \sum_{m=1}^{n_2} |X_i - Y_m| ,
\]

\[
- \frac{2}{n^2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_1} |X_i - X_j| ,
\]

\[
- \frac{2}{n^2} \sum_{i=1}^{n_1} \sum_{m=1}^{n_2} |Y_i - Y_m|. \quad (4)
\]

For two random vector \(X = X_1, ..., X_n\), and \(Y = Y_1, ..., Y_n\), the energy distance \(\varepsilon(X, Y)\) is defined as in [15].

For multiple random vector \(X_i\), the energy \(E_m\) is defined as follows [15]:

\[
E_m(X_i) = \sum_{1 \leq j < k < K} \left( \frac{n_j + n_k}{2N} \right) \left( \frac{n_j n_k}{n_j + n_k} \varepsilon(X_j, X_k) \right). \quad (5)
\]

### 2.3 SVM-RFE-ED Algorithm

The proposed approach is an enhanced version of standard SVM-RFE and it incorporates the energy distance to compute the class separability. SVM-RFE-ED uses a new modified rank score defined as follows:

**Algorithm 2 SVM-RFE-ED Algorithm**

1. Initialisation of \(\sigma\)
2. Given set of genes, \(G = \{X_1, ..., X_n\}\)
3. Initialisation \(S = G\)
4. Ranked set of genes, \(R = \{\}\)
5. Select the type of kernel function
6. Initialisation of parameters \(C\) and kernel parameters
7. Initialisation of \(\beta\)
8. repeat
9. Train SVM on \(S\)
10. for \(X_i \in S\) do
11. Compute \(w_i\)
12. end for
13. Select the genes \(X_i^*\) with the smallest ranking \(w_i < \sigma\)
14. Update \(R = R \cup \{X_i^*\}\) and \(S = S \setminus \{X_i^*\}\)
15. until All genes are ranked
16. Output: Rank list according to smallest score, \(R\)

\[
e_i = \beta \times w_i + (1 - \beta) \times E_m, \quad (6)
\]

where \(e_i\) is the rank score of the \(i\)th gene, \(\beta\) is a parameter that determine the tradeoff between SVM weights and energy distance.

The algorithm of SVM-RFE-ED is described as follows:

To demonstrate the performance of Energy Distance we propose to compare SVM-RFE using energy distance with SVM-RFE using Hausdorff distance and Jeffries-Matusia (JM) distance.
**Hausdorff distance** is developed by Nadler in 1978 [12, 13] and it computes the similarity between two vectors. The basic idea of Hausdorff distance assumes that two groups are similar. For two groups \(X\) and \(Y\), the Hausdorff distance \(D_H(X, Y)\) is defined as follows:

\[
D_H(X, Y) = \max\{h(X, Y), h(Y, X)\},
\]

(7)

\[
h(X, Y) = \max_{x_i \in X} \min_{x_j \in Y} \|x_i - x_j\|,
\]

(8)

The function \(h(X, Y)\) is called the direct Hausdorff distance from \(X\) to \(Y\) [5].

**Jeffries-Matusia (JM) distance** is widely used for variable selection [1, 4]. For two vectors \(X\) and \(Y\), the JM distance is defined as follows:

\[
D_{JM}(X, Y) = \sqrt{2(1 - e^B)},
\]

(9)

\[
B(X, Y) = \frac{1}{8} (\mu_X - \mu_Y)^T \left( \frac{\Sigma_X + \Sigma_Y}{2} \right) (\mu_X - \mu_Y) + \frac{1}{2} \ln \left[ \frac{\Sigma_X + \Sigma_Y}{\Sigma_X^2 \Sigma_Y^2} \right],
\]

(10)

where \(\mu\) is the class mean vector and \(\Sigma\) the class covariance [1].

**Table 2.** Information about the microarray datasets

<table>
<thead>
<tr>
<th>Dataset name</th>
<th>Number of genes</th>
<th>Number of samples</th>
<th>Number of classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Colon</td>
<td>2000</td>
<td>62</td>
<td>2</td>
</tr>
<tr>
<td>DLBC</td>
<td>4026</td>
<td>47</td>
<td>2</td>
</tr>
<tr>
<td>Leukemia</td>
<td>5147</td>
<td>72</td>
<td>2</td>
</tr>
<tr>
<td>Lung</td>
<td>12533</td>
<td>181</td>
<td>2</td>
</tr>
<tr>
<td>Ovarian</td>
<td>15154</td>
<td>253</td>
<td>2</td>
</tr>
</tbody>
</table>

**3 Experimental Results**

**3.1 Datasets**

In this section, we present the results obtained by the proposed approach. The experimentations are conducted on five datasets widely used to benchmark gene selection approaches, namely, colon cancer, leukemia cancer, lung cancer, ovarian cancer and DLBC cancer. Table 2 presents the information about the datasets used in our study.

The first column of table 2 presents the name of dataset. The second column is the number of genes and the last column contains the number of samples.

**3.2 Parameters Setting**

We randomly split the original dataset into separate training and testing sets. Table 3 shows the number of genes and samples used for training and testing phase in each dataset.

**Table 3.** Number of samples used for training and testing

<table>
<thead>
<tr>
<th>Dataset name</th>
<th>Training</th>
<th>Testing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Colon</td>
<td>40</td>
<td>22</td>
</tr>
<tr>
<td>DLBC</td>
<td>28</td>
<td>19</td>
</tr>
<tr>
<td>Leukemia</td>
<td>43</td>
<td>29</td>
</tr>
<tr>
<td>Lung</td>
<td>108</td>
<td>73</td>
</tr>
<tr>
<td>Ovarian</td>
<td>151</td>
<td>102</td>
</tr>
</tbody>
</table>

The first column of table 3 presents the name of dataset. The second column is the number of samples used for training and the last column contains the number of samples used for the test.

The algorithm of SVM-RFE-ED is trained by using a kernel function. In this work, we propose to use four kernel functions. Table 4 presents the information about the parameters setting of each kernel function.

The first column of table 4 presents the name of the kernel function and the second column is the value of the kernel parameters. These parameters
have been chosen by experimentation and have proven their performance.
The parameter $C$ of the SVM-RFE-ED algorithm is set to 1. The value of $\beta$ using to compute the equation (6) is set to 0.5.

### 3.3 Results and Discussions

#### Table 5. Classification accuracy rate (CAR), Sensitivity and Specificity obtained by the proposed approach for each dataset

<table>
<thead>
<tr>
<th>Datasets</th>
<th>CAR</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Colon</td>
<td>95.65</td>
<td>0.93</td>
<td>1</td>
</tr>
<tr>
<td>DLBC</td>
<td>100</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Leukemia</td>
<td>100</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Lung</td>
<td>100</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Ovarian</td>
<td>100</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

The performance evaluation of the proposed approach SVM-RFE-ED is conducted in terms of: classification accuracy rate, sensitivity and specificity. Table 5 and 6 show the performance and the results obtained by the proposed approach.

Table 5 gives the classification accuracy rate (CAR), sensitivity and specificity of our approach for each dataset. As seen, the performance of SVM-RFE-ED was significantly better. The proposed approach has reached 100% of the classification accuracy rate and significantly improved the sensitivity and specificity for DLBC, leukemia, lung and ovarian datasets.

Table 6 presents the number of selected genes and the kernel functions that have provided better results. The analysis of the results show that the proposed approach provided better results with respect to the number of selected genes. As seen, we remark that the number of selected has been significantly reduced. For DLBC, leukemia, lung and ovarian datasets the best accuracy is recorded for the 5% of selected genes i.e. after ranking the genes, the 5% of genes that have the high score have given the better results. For colon cancer dataset, the high classification accuracy rate is obtained by the 30% of genes. Compared to initial number of genes, the proposed approach has largely reduced the number of genes.

The second column of table 6 describes the best kernel functions that have provided very good results. In colon cancer, the polynomial kernel has provided good results. For DLBC, the multilayer perceptron kernel has given good results. In leukemia cancer, the linear, Gaussian, polynomial and multilayer perceptron kernel provided a 100% accuracy. In lung cancer, the kernels: linear, polynomial and quadratic achieved a 100% of accuracy. For ovarian cancer, linear, Gaussian, polynomial and quadratic kernels provided better results.

The results obtained by the proposed approach SVM-RFE-ED are summarized on the following figures.

The figures 1, 3, 2, 4 and 5 illustrate the classification accuracy rate obtained by the proposed approach for each dataset and for some percentage of selected genes. We compute the classification accuracy rate for 5% of genes to 100% of genes. We clearly show that the best results is obtained for 5% of genes. The class separability measures combined with the weight vector generated by SVM improve significantly the...
The proposed approach SVM-RFE-ED uses the energy distance to compute the class separability for each gene. To test the performance and the results produced by SVM-RFE using energy distance, we propose to use two other distances Hausdorff distance and Jeffries-Matusia (JM). The results are described on table 7.

The results illustrated on table 7 show that the classification accuracy rate obtained by using Hausdorff and JM distances are slightly identical compared to SVM-RFE-ED. We observe a small advantage for SVM-RFE-ED in lung and ovarian datasets.

To validate the performance and the results obtained by the proposed approach SVM-RFE-ED, we propose to compare the results of classification performances obtained by our approach with the results of seven gene selection approaches reported from [11]. Table 8 and figure 6 describes these results.

Table 8 shows the results of classification accuracy rate obtained by SVM-RFE-ED and compared to seven approaches of gene selection. The first column of table 8 represents the name of gene selection approaches. The second and the third columns are the classification accuracy rate of colon cancer and leukemia respectively.

The analysis of the results of table 8 demonstrates that the proposed approach SVM-RFE-ED provides satisfactory results and achieves a high classification accuracy rate compared to other approach. As seen, the classification accuracy rate and reduce largely the number of selected genes.
Fig. 5. Classification accuracy rate for each selected genes obtained by using SVM-RFE-ED with Multilayer Perceptron Kernel

Table 7. Classification accuracy rate (CAR), obtained by the proposed approach SVM-RFE-ED and SVM-RFE using Hausdorff and Jeffries-Matusia distances

<table>
<thead>
<tr>
<th>Datasets</th>
<th>SVM-RFE-ED</th>
<th>Hausdorff</th>
<th>JM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Colon</td>
<td>95.65</td>
<td>94.66</td>
<td>95.50</td>
</tr>
<tr>
<td>DLBC</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Leukemia</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Lung</td>
<td>100</td>
<td>99.95</td>
<td>100</td>
</tr>
<tr>
<td>Ovarian</td>
<td>100</td>
<td>99.90</td>
<td>99.98</td>
</tr>
</tbody>
</table>

performances are significantly better for the both datasets cancer and leukemia.

In order to validate the results and the performances of the proposed approach SVM-RFE-ED, we must measure its stability. The stability of feature selection method is defined as the sensitivity of a method to variations in the training set, in other term, the stability is the measure or robustness of a method when the training set is different [7]. In this study, we compute two stability measure widely used in the literature: $S_S$ and $S_H$.

$S_S$ stability was proposed by Kalousis et al. [7] and it is defined as follows:

$$S_S = \frac{|A \cap B|}{|A \cup B|}. \quad (11)$$

$S_H$ stability was developed by Dunne et al. [3] and is defined as the relative Hamming distance. This measure is defined as follows:

$$S_H = 1 - \frac{|A \setminus B| + |B \setminus A|}{n}, \quad (12)$$

where:

- $A$ and $B$ are a set of selected features using different training set,
- $n$ is the total number of features.
- $\setminus$ is the set-minus.

The values of $S_S$ and $S_H$ are between $[0, 1]$. We can compute the stability for many subset of selected features by computing the average of all pairwise.

In this study, we run the proposed approach SVM-RFE-ED 20 times by using 20 different training sets. The results are illustrated in figures 7 and 8.
Figures 7 and 8 show the ss stability and sh stability computed for each dataset by using 20 different training sets. Each blue box in the figures indicates the upper and lower quartiles. The small circle indicates the median value. As seen in figures 7 and 8 the lower value of $S_S$ stability and $S_H$ stability for each dataset are between 0.85 and 0.9, the upper value are between 0.96 and 1. These values of stability are very close to 1 which means that the proposed approach SVM-RFE-ED is very robust and produces a stable subsets of features if we change the training set.

4 Conclusion

In this paper, we address the problem of cancer diagnosis by solving the gene selection problem. We propose a novel SVM-RFE based on energy distance. The proposed approach was called SVM-RFE-ED and it combines the weight vector provided by SVM and the energy distance to measure the class separability of each gene. The performance evaluation has been conducted on five widely used datasets of cancer diagnosis: colon, DLBC, leukemia, lung and ovarian.

Though the results obtained by the proposed approach, we have clearly observed that SVM-RFE-ED provided very good results by reducing significantly the number of genes. In addition, the stability of SVM-RFE-ED has been demonstrated. Hence, in future we would considering the problem of genes redundancy and incorporate this problem on SVM-RFE.

References


Article received on 28/10/2017; accepted on 17/05/2018. Corresponding author is Seyyid Ahmed Medjahed.