ISSN 2007-9737

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

Seyyid Ahmed Medjahed[1,2], Mohammed Ouali[3]

[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 approched 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):

$$\begin{array}{ll} \min_{w,b,\xi} & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\xi_i, \\ subject\ to & y_i(\langle w, x_i\rangle + b) \geq 1 - \xi_i, \\ & \xi_i \geq 0, \\ & i \in \{1, ..., N\}, \end{array} \quad (1)$$

where ξ_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:

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

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 α_i is the solution of the dual problem. The primal solution is given as follows:

$$\begin{aligned} w^* &= \sum_{i=1}^{N}\alpha_i y_i x_i, \quad (3) \\ b^* &= -\frac{1}{2}\langle w^*, x_r + x_s\rangle . \end{aligned}$$

Table 1. The most used kernel functions

Kernel name	Formulation	Parameters
Linear	$k(x, y) = x^t.y$	/
Polynomial	$k(x, y) = (x^t.y)^d$	d
Gaussian	$k(x, y) = \exp(-\frac{\|x-y\|^2}{2\sigma^2})$	σ
Multilayer Perceptron	$k(x, y) = \tan(P1.x^t.y + P2)$	P_1, P_2
Quadratic	$k(x, y) = (x^t.y + 1)^2$	/

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

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

α_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

Algorithm 1 SVM-RFE Algorithm

1: Initialisation of σ
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: **repeat**
8: Train SVM on S
9: **for** $X_i \in S$ **do**
10: Compute w_i
11: **end for**
12: Select the genes X_i^* with the smallest ranking $w_i < \sigma$
13: Update $R = R \cup \{X_i^*\}$ and $S = S \setminus \{X_i^*\}$
14: **until** All genes are ranked
15: Output: Rank list according to smallest score, R

into consideration the class separability of each gene. In order to overcome this limitation, we propose to improve SVM-RFE by incorporates the energy distance that computes the measure of discrimination of each gene.

2.2 Energy Distance

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]:

$$\begin{aligned}\varepsilon(X,Y) &= \frac{2}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{m=1}^{n_2}|X_i - Y_m|, \\ &\quad -\frac{2}{n_1^2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_1}|X_i - X_j|, \\ &\quad -\frac{2}{n_2^2}\sum_{l=1}^{n_2}\sum_{m=1}^{n_2}|Y_l - Y_m|. \end{aligned} \quad (4)$$

For two random vector $X = X1, ..., X_{n_1}$ and $Y = Y1, ..., Y_{n_2}$, 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]:

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

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 σ
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 β
8: **repeat**
9: Train SVM on S
10: **for** $X_i \in S$ **do**
11: Compute e_i using equation (6)
12: **end for**
13: Select the genes X_i^* with the smallest ranking $e_i < \sigma$
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

$$e_i = \beta \times w_i + (1 - \beta) \times E_{m_i}, \quad (6)$$

where e_i is the rank score of the ith gene, β 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:

$$D_H(X,Y) = \max\{h(X,Y), h(Y,X)\}, \quad (7)$$

$$h(X,Y) = \max_{x_i \in X} \min_{x_j \in Y} \|x_i - x_j\|, \quad (8)$$

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:

$$D_{JM}(X,Y) = \sqrt{2(1-e^B)}, \quad (9)$$

$$B(X,Y) = \frac{1}{8}(\mu_X - \mu_Y)^T \left(\frac{\Sigma_X + \Sigma_Y}{2}\right)(\mu_X - \mu_Y) + \frac{1}{2}\ln\left[\frac{\left|\frac{\Sigma_X+\Sigma_Y}{2}\right|}{|\Sigma_X|^{\frac{1}{2}}|\Sigma_Y|^{\frac{1}{2}}}\right], \quad (10)$$

where μ is the class mean vector and Σ the class covariance [1].

Table 2. Information about the microarray datasets

	Number of		
Dataset name	genes	samples	classes
Colon	2000	62	2
DLBC	4026	47	2
Leukemia	5147	72	2
Lung	12533	181	2
Ovarian	15154	253	2

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 colon 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

Dataset name	Training	Testing
Colon	40	22
DLBC	28	19
Leukemia	43	29
Lung	108	73
Ovarian	151	102

The first column of table 3 presents the name of dataset. The second colon 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

Table 4. Kernel function and parameters setting

Kernel name	Parameters setting
Linear	/
Polynomial	$d = 3$
Gaussian	$\sigma = 0.0156$
Multilayer Perceptron	$P_1 = 0.5\ P_2 = 1$
Quadratic	/

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 β 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

Datasets	CAR	Sensitivity	Specificity
Colon	95,65	0,93	1
DLBC	100	1	1
Leukemia	100	1	1
Lung	100	1	1
Ovarian	100	1	1

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.

Table 6. The number of selected genes and the best kernel for each dataset

Datasets	Selected Genes	Kernel
Colon	600	Polynomial
DLBC	201	Multilayer Perceptron
Leukemia	257	Linear, Gaussian Polynomial, Multilayer Perceptron
Lung	626	Linear, Polynomial, Quadratic
Ovarian	757	Linear, Gaussian, Polynomial, Quadratic

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

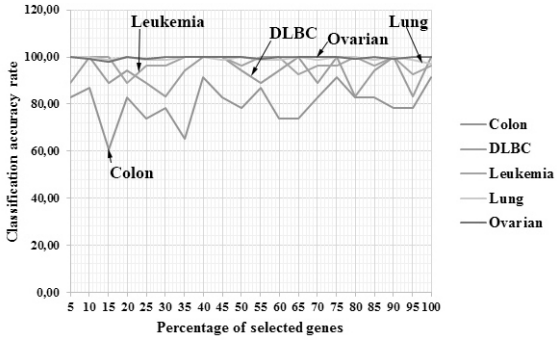

Fig. 1. Classification accuracy rate for each selected genes obtained by using SVM-RFE-ED with Linear Kernel

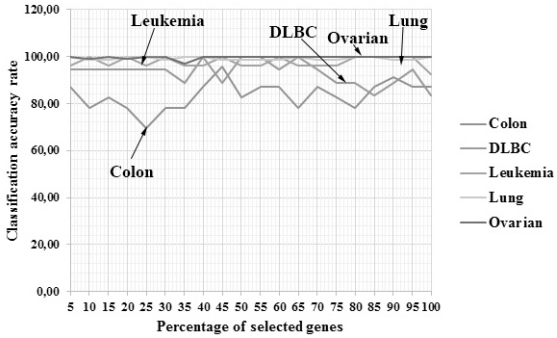

Fig. 2. Classification accuracy rate for each selected genes obtained by using SVM-RFE-ED with Quadratic Kernel

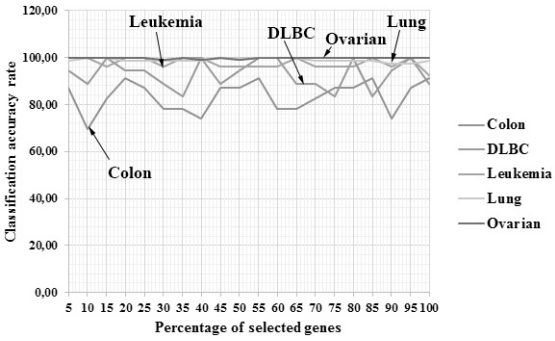

Fig. 3. Classification accuracy rate for each selected genes obtained by using SVM-RFE-ED with Gaussian Kernel

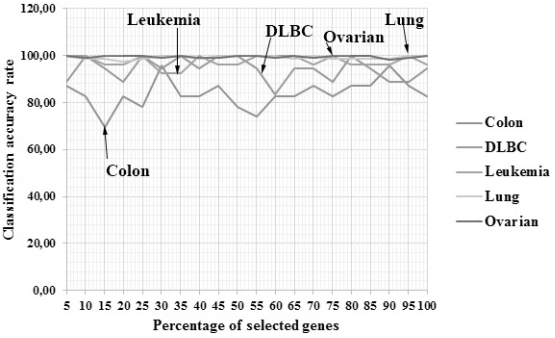

Fig. 4. Classification accuracy rate for each selected genes obtained by using SVM-RFE-ED with Polynomial Kernel

classification accuracy rate and reduce largely the number of selected genes.

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

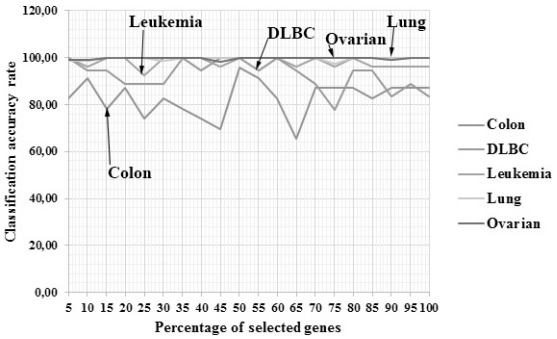

Fig. 5. Classification accuracy rate for each selected genes obtained by using SVM-RFE-ED with Multilayer Perceptron Kernel

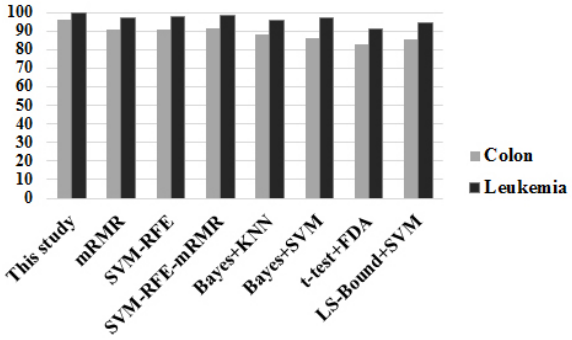

Fig. 6. Classification accuracy rate obtained by the proposed approach SVM-RFE-ED compared to other approaches

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

		SVM-RFE	
Datasets	SVM-RFE-ED	Hausdorff	JM
Colon	95,65	94,66	95,50
DLBC	100	100	100
Leukemia	100	100	100
Lung	100	99,95	100
Ovarian	100	99,90	99,98

Table 8. Comparison of gene selection approach with the proposed approach SVM-RFE-ED

Methods	Colon	Leukemia
This study	95,65	100
mRMR [11]	91,00	97,18
SVM-RFE [11]	91,00	97,88
SVM-RFE-mRMR [11]	91,68	98,38
Bayes + KNN [11]	88,23	95,71
Bayes + SVM [11]	86,27	97,12
t-test + FDA [11]	82,68	90,86
LS-Bound + SVM [11]	85,23	94,74

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:

$$S_S = \frac{|A \cap B|}{|A \cup B|}. \tag{11}$$

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

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

where:

A and B are a set of selected features using different training set,

n is the total number of features.

$|.|$ is the cardinality,

'$\setminus$' 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.

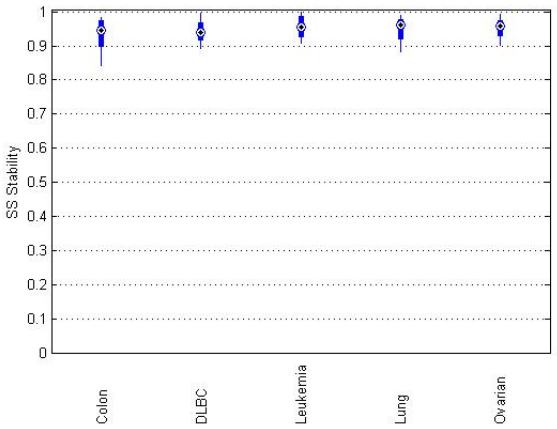

Fig. 7. S_S stability computed for each dataset

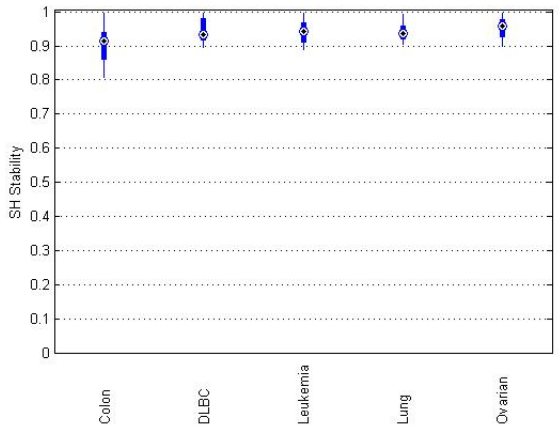

Fig. 8. S_H stability computed for each dataset

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

1. **Bruzzone, L., Roli, F., & Serpico, S. B. (1995).** An extension of the Jeffreys-Matusita distance to multiclass cases for feature selection. *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 33, No. 6, pp. 1318–1321.
2. **Cortes, C. & Vapnik, V. (1995).** Support-vector networks. *Machine Learning*, Vol. 20, No. 3, pp. 273–297.
3. **Dunne, K., Cunningham, P., & Azuaje, F. (2002).** Solution to instability problems with sequential wrapper-based approaches to feature selection. *Technical Report, Journal of Machine Learning Research*.
4. **Hao, P., Zhan, Y., Wang, L., Niu, Z., & Shakir, M. (1995).** Feature selection of time series MODIS data for early crop classification using random forest: A case study in Kansas, USA. *Remote Sensing*, Vol. 7, No. 5, pp. 5347–5369.

5. **Huttenlocher, D. P., Klanderman, G. A., & Rucklidge, W. J. (1993).** Comparing images using the Hausdroff distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 15, No. 9, pp. 580–863.
6. **Isabelle, G., Jason, W., Stephen, B., & Vladimir, V. (2002).** Gene selection for cancer classification using support vector machines. *Machine Learning*, Vol. 46, No. 1, pp. 389–422.
7. **Kalousis, A., Prados, J., & Hilario, M. (2007).** Stability of feature selection algorithms: a study on high-dimensional spaces. *Knowledge and Information Systems*, Vol. 12, No. 1, pp. 95–116.
8. **Medjahed, S. A., Saadi, T. A., Benyettou, A., & Ouali, M. (2016).** Microcanonical annealing and threshold accepting for parameter determination and feature selection of support vector machines. *CIT Journal of Computing and Information Technology*, Vol. 24, No. 4, pp. 369–382.
9. **Mishra, S. & Mishra, D. (2015).** Svm-bt-rfe: An improved gene selection framework using bayesian t-test embedded in support vector machine(recursive feature elimination) algorithm. *Karbala International Journal of Modern Science*, Vol. 1, No. 2, pp. 86–96.
10. **Mohankumar, S. & Balasubramanian, V. (2016).** Identifying effective features and classifiers for short term rainfall forecast using rough sets maximum frequency weighted feature reduction technique. *CIT Journal of Computing and Information Technology*, Vol. 24, No. 2, pp. 181–194.
11. **Mundra, P. A. & Rajapakse, J. C. (2010).** SVM-RFE with MRMR filter for gene selection. *IEEE Transactions On Nanobioscience*, Vol. 9, No. 1, pp. 31–37.
12. **Nadler, S. B. (1979).** Hyperspaces of sets. *Bulletin (New Series) of the American Mathematical Society*, Vol. 1, No. 2, pp. 412–414.
13. **Piramuthu, S. (1999).** The Hausdorff distance measure for feature selection in learning applications. *Proceedings of the 32nd Hawaii International Conference on System Sciences*.
14. **Quintanilla-Domínguez, J., Ruiz-Pinales, J., Barrón-Adame, J. M., & Cabrera, R. G. (2018).** Microcalcifications detection using image processing. *Computación y Sistemas*, Vol. 22, No. 1, pp. 291–300.
15. **Szekely, G. J. & Rizzo, M. L. (2013).** Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, Vol. 143, No. 8, pp. 1249–1272.
16. **Waheeb, W. & Ghazali, R. (2017).** Content-based sms classification: Statistical analysis for the relationship between features size and classification performance. *Computación y Sistemas*, Vol. 21, No. 4, pp. 771–785.
17. **Zhou, Y., Xu, J., Cao, J., Xu, B., Li, C., & Xu, B. (2017).** Hybrid attention networks for Chinese short text classification. *Computación y Sistemas*, Vol. 21, No. 4, pp. 759–769.

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