Journal of Engineering Technology
Volume 6, Issue 2, July. 2018, PP. 98-113
RESEARCH ARTICLE -ENGINEERNG TECHNOLOGY

New approaches for gene selection and cancer diagnosis based on microarray gene expression profiling

Sara Haddou Bouazza[1], Khalid Auhmani[2] , Abdelouhab Zeroual[1]

[1]Department of Physics, Faculty of Sciences Semlalia, Cadi Ayyad University, Marrakech, Morocco
[2]Department of Industrial Engineering, National School of Applied Sciences, Cadi Ayyad, Safi, Morocco

Abstract: Selecting a suitable subset of informative genes became an important task to analyze cancer's data sets. For this reason, we employed a feature selection technique to reduce the dimensionality of the DNA Microarray data. This operation aims to select a relevant subset of genes from the data set without any transformation of the original data.
This paper proposes two new approaches for cancer classification. The first approach is a Gene selection method based on three filters: The first filter is a direct use of a filter selection method, the second filter searches for the highest accuracy by removing noisy genes, and the third filter reduces the dimensionality of the selected genes by selecting a minimum subset of relevant genes. The second approach is a Gene classification method, composed of two steps: The first step is based on the Statistic measures of the selected genes, and the second step compares the vote of genes. We named the Gene selection method by the Three Filter selection method, and the Gene classifier by the Statistic classifier.
We performed our approaches on three cancers: Leukemia, Colon cancer, and Prostate cancer. We used three filter selection methods: signal to noise ratio, correlation coefficient, and ReliefF. We applied three classifiers: K Nearest Neighbor, Support Vector Machine, and Linear Discriminant Analysis. The obtained results showed that, genes selected by our Three Filter selection method provide the highest accuracies for the smallest subset of relevant genes. Also, our Statistic classifier is simpler, faster, and provides the highest accuracies compared to well-known classic classifiers.

Keywords: Feature selection; supervised classification; image processing; DNA Microarray.

1 Introduction

A gene is a segment of DNA Microarray which contains all the information necessary to create all sorts of proteins in our body. It measures expression levels of thousands of genes simultaneously [1]. This technology has proven to be encouraging in predicting cancer classification and prognosis outcomes [2].

The Major issue in the DNA Microarray technology is the fact that it present expression levels of thousands of genes for only a dozen of samples. In this case we need to reduce the high dimensionality of genes, by removing redundant genes and keeping informative and relevant genes.
One way to resolve this issue is Feature subset selection, especially for the Gene Classification area.

In Microarray Gene classification, feature selection and classification are two main problems. They are related in the sense that, when a good set of features is identified, the classification model

yields high accu-racy. Thus, the Microarray Gene classification is a supervised learning task which uses gene expression array phenotype to predict the diagnosis of a sample. It generates a classify model, from labelled gene expression data samples, to classify new data samples into different predefined diseases. The major challenge in Gene Classification is feature selection [3].

The Feature selection process is the removal of redundant and irrelevant features. The latter reduces the di-mensionality of the data to be processed by the classifier and improves the prediction accuracy [4].

In this paper, we propose two new approaches, the first for Gene selection and the second for Gene classification.

The first approach is a Three filter selection method. This new method is based on the known filter selection approaches. Its first filter is a direct application of a classic filter approach selection method, its second filter removes noisy genes, and its third filter removes redundancy genes.

The second approach is a Statistic Classifier based on gene expression statistical measures. It classifies a sample based on genes's statistical information on all the existed classes. Each gene proposes a class of the sample data, and the classifier decides the final class by taking the most voted class by the majority of genes.

2. Statistics-Based Gene selection methods

With the presence of a large number of features and a limited number of samples, a learning model tends to overfit; Consequently, resulting in their performance degenerates [5]. In order to resolve this problem, dimensionality reduction techniques became an obligatory task in the machine learning and data-mining research area [6].

For Gene classification, the most used gene selection methods are based on gene ranking. Each gene is analyzed individually and assigned a score corresponding to its correlation with the class. Genes are then ranked by their scores and the top-ranked ones are selected as relevant. Gene selection methods can be divided into two methods: wrapper and filter methods [7], [8].

A filter method performs gene selection independently to preprocessing of the Microarray dataset. This method has the advantage of reducing the data set size before classification. However, it does not take the relationships between genes into account. Some selected genes have similar expression levels and this redundancy affects the performances of classification; Whereas, a wrapper method embeds a gene selection method within a classification algorithm. Kohavi and John [7] have discovered that the wrapper methods can improve the accuracy of classification algorithms over the filter methods.

In order to deal with gene expression Microarray data more effectively and efficiently, classification algorithms need to consider applying a combination of filter and wrapper methods for Microarray data selection. In this work, we study three parametric methods: Signal-to-noise ratio (SNR), Correlation Coefficient (CC), and ReliefF.

2.1 Signal-to-noise ratio (SNR)

The signal to noise ratio (SNR) method identifies the expression patterns with a maximal difference in mean expression between two groups and minimal variation of expression within each group [9], [10].

This criterion was proposed by [11] and sets the score as follows:

$$P(j) = \frac{M1j - M2j}{S1j + S2j} \quad (1)$$

Where M_{kj}, S_{kj} denotes the mean and the standard deviation of the gene j for samples of classes k = 1, 2. Higher values of the score |P (j) | indicate a strong correlation between the values of the gene and the distinction of classes.

2.2 Correlation Coefficient (CC)

Correlation coefficients measure the strength of association between two genes. The most common correlation coefficient, called the Pearson product-moment correlation coefficient, measures the strength of the linear association between genes [12].

The sign and the absolute value of a Pearson correlation coefficient describe the direction and the magnitude of the relationship between two genes.

- The value of a correlation coefficient ranges between -1 and 1.
- The higher the absolute value of a correlation coefficient, the stronger is the linear relationship.
- The strongest linear relationship is indicated by a correlation coefficient of -1 or 1.
- The weakest linear relationship is indicated by a correlation coefficient equal to 0.
- A positive correlation means that if one gene gets bigger, the other gene tends to get more bigger.
- A negative correlation means that if one gene gets bigger, the other gene tends to get smaller.

A formula for computing a Pearson correlation coefficient is given below. So if we consider a data set {X1,...,Xn} containing n samples with corresponding classes {Y1,...,Yn} where each sample Xi is composed of N genes, then the formula r_j for a gene j is:

- $$r_j = \frac{\sum_{i=1}^{n}(Xij - \bar{X}j)(Yi - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(Xij - \bar{X}j)^2}\sqrt{\sum_{i=1}^{n}(Yi - \bar{Y})^2}} \quad (2)$$

Where:

- r denotes the score of the Pearson correlation coefficient
- Xij is the i^{th} sample value for the gene j and Yi is the corresponding class
- $\bar{X}j = \frac{1}{n}\sum_{i=1}^{n} Xij$ (the sample mean for a gene j); and analogously for $\bar{Y}$

2.3 ReliefF

ReliefF is the supervised feature weighting algorithms of the filter approach [13]. It is introduced as Relief by [14] and then improved by Kononenko as the ReliefF [15].

The vector W of estimations of the qualities of attributes:

$$W_A = w_d - \sum_{j=1}^{k} \frac{\text{diff (Ai,Xi,hits j)}}{m*k} + \sum_{c \neq class(Xi)} \frac{p(c)}{1 - p(class(Xi))} \sum_{j=1}^{k} \frac{diff(Ai,Xi,missesj)}{m*k} \quad (3)$$

The distance used is defined by:

$$diff(Ai, X1, X2) = \frac{|value(Ai,X1) - value(Ai,X2)|}{\max(A) - \min(A)} \quad (4)$$

Xi is an instance described by the vector Ai of n genes; m is the number of repetitions of the process, it is a user-defined parameter; k is the number of the nearest misses; hits j is the nearest hit, misses j is the nearest miss; diff (Ai, X1, X2) calculates differences between the values of the attribute Ai for two instances X1 and *X2*.

A way to improve prediction classification accuracy, as well as interpretation of the relationship between genes and the considered cancer, is to apply a supervised classification task based on expression values of the identified genes selected by an effective gene selection method.

3 Gene Classification methods

Machine learning is the subfield of computer science which allowed computers to learn without being programmed [12]. Prediction is the main task used in many research areas, including machine learning, pattern recognition, signal and image processing, and research information, etc. [16]

Supervised Gene classification is the process of discriminating data, a set of samples, so that the objects in the same group (called classes) are closer (as a criterion of similarity) to each other than other groups. [17].

We evaluated the Feature Selection methods by calculating the classification accuracy of the three classifiers: K nearest neighbors (KNN), Support Vector Machines (SVM), and Linear Discriminant Analysis (LDA).

3.1 K-Nearest Neighbor

The k-nearest neighbor classifier is one of the simplest and commonest used of all machine learning methods.

K nearest neighbor classifier (KNN) is based on the principal of close proximity evaluated by calculating the Euclidean distance between the test sample and the training samples [18].

3.2 Support Vector Machine

Support Vector Machine (SVM) is a supervised learning classifier which uses a kernel function to implicitly map data in a high dimensional space. Then, by solving an optimization problem with the training data, it constructs the maximum margin hyper plane [19], [20].

3.3 Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) is a generalization of Fisher's linear discriminant, an algorithm used in the machine learning to search and find a linear combination of features which characterizes or separates two or more classes of objects.

Linear Discriminant Analysis easily handles the case where the within-class frequencies are unequal and their performance has been examined on randomly generated test data. This method maximizes the ratio of between-class variance of the within-class variance in any particular data set thereby guaranteeing maximal separability [21].

To evaluate the performances of the classifiers, we measure the value of the classification accuracy [22]:

$$\text{Accuracy} = 100^* (TP + TN) / (TN + TP + FN + FP) \quad (5)$$

Where TP is true positive for correct prediction to disease class, TN is true negative for correct prediction to normal class, FP is false positive for incorrect prediction to disease class, and FN is false negative for incorrect prediction to normal class.

4 A new approach for Gene selection for binary class problems

Gene selection is a principal process in Gene classification. The main problem in a particular classification task is the identification of genes expressed differently between the various classes, since most genes are irrelevant and uninformative to the given class. For a selection task, the problem of gene selection is to find among the entire high dimensional space of genes, a subset that best characterizes the response target variable.

It exists many ways to select a relevant subset of genes: Filters methods and wrapper methods.

The filter methods attempt to evaluate the importance of each gene statically according to a heuristic scoring criterion independently of any particular classifier [23]. The highest scoring genes are selected and applied to a classification process.

The researcher Marczyk et al. [24] suggest the use of adaptive filter methods based on the decomposition of the probability density function of gene expression means or variances into a mixture of Gaussian components. M. Dashtban et al. [25] proposed a method based on genetic algorithms and artificial intelligence to identify predictive genes for cancer classification. A filter method was first applied to reduce the dimensionality of the feature space, followed by employing an integer-coded genetic algorithm with dynamic-length genotype, intelligent parameter settings, and modified operators. M. Dashtban et al. [26] presented a hybrid model using the Fisher criterion applied to three widely-used microarray cancer datasets. Haddou Bouazza. S et al. [27] propose the use of the scoring selection methods Fisher, T-Statistics, SNR and ReliefF. They suggest the use of K Nearest Neighbors and Support Vector Machines as supervised classifiers. Osama Mahmoud et al. [28] propose a statistical method to detect relevant genes based on overlapping analysis of genes expression values across classes. The method analyzes the expressions overlap across target classes. Laussen B et al. [29] propose an Adaptive filtering of microarray gene expression data based on Gaussian mixture decomposition, which allows determining close to optimal threshold values for sample means and sample variances for gene filtering. Haddou Bouazza. S [30] suggest the use of Correlation Coefficient and Max-relevance Min-Redundancy to select relevant genes, then classified by the supervised classifier K Nearest Neighbors, Support Vector Machines, Linear Discriminant Analysis and Decision Tree for supervised classification. Ultsch A et al. [31] propose an algorithm, called 'PUL', in which the differentially expressed genes are identified based on a measure of retrieval information named PUL score. Lu et al. [32] suggest a method to select relevant genes in which principle component analysis has been used to discover the sources of variation in genes expression values and to remove genes corresponding to components with less variation.

The major problems is that the filter methods focus on the utility of individual gene only and ignore the combination of genes, and also, the optimal size of genes subset is hard to be determined.

On the other hand, the wrapper methods search relevant genes through using a classifier to measure the significance of a candidate gene subset [33], [34].

Hybrid methods between the filter and wrapper methods use the filter for the pre-selection, and then the wrapper for determinating the optimal size of genes and for getting high accuracy. Considering the overlap between the identified gene expression values; for different classes; turns to be another necessary criterion to recognize significant genes and relevant to the classification task.

This operation utilities the information given by sample classes and gene expression values, to recognize a smaller subset of expressed genes between target classes. A classifier is used to select those genes to improve the classification performance and prediction accuracy of the selected subset of genes.

This strategy allows the detection of a reduced set of relevant genes which gives the best classification coverage on training samples.

Shamsul [35] suggested a combination between the filter and the wrapper approach which combine between a Mutual Information based on Maximum Relevance and a wrapper based on Artificial Neural Network. In the same way, Cadenas [36] presented a method that uses a Fuzzy Random Forest, the authors integrate filter and wrapper approaches with a sequential search strategy.

Li-Yeh Chuang [37] applied the information gain selection method as a filter approach, and an improved binary particle swarm optimization as a wrapper approach to implement feature selection. The authors compare gene selection performances of the filter and wrapper methods, and hybrid the two approaches to produce a hybrid model for gene selection. They utilized K-nearest neighbor and a Support Vector Machine classification to evaluate the performances of the selected subset of genes.

The main differences between our approach and [35], [36], [37] is the use of Signal-to-noise ratio, Correlation Coefficient and ReliefF as filter methods combined with a Search for the Highest Accuracy as a wrapper strategy, with the use of K nearest neighbors, Support Vector Machines, Linear Discriminant Analysis, Decision Tree for Classification and Naïve Bayes supervised classifier. As well as a third step which selects the minimum subset of genes. This subset is designated to be the minimum one that correctly classifies the maximum number of samples in a given training set. Such a procedure permits removing redundant genes with similar expression profiles.

Selecting a minimum subset of genes is a step in which the information provided by the selection method is analyzed.

Baralis et al. [38] have proposed a procedure to detect a minimum subset of genes. The major differences are that [38] use the expression range to define the intervals which are employed for constructing gene masks, and then apply a set covering approach to obtain the minimum feature subset. The same technique is performed by [39] to get a minimum gene subset using a greedy approach rather than the set covering.

The main idea of this paper is to present a new effective selection method which uses a combination between the filter and the wrapper methods to select the minimum subset of relevant and informative genes.

Our new selection method is based on three steps (Figure 1):

The first step is an application of a Filter selection method on the entire data set. As more genes are included, the accuracy of the classification scheme varies.

The filter selection methods permit to rank each gene and select a subset of relevant genes. The problem is that the existed noisy genes affect on the other selected genes, and decrease the classification accuracy.

In order to exclude those noisy selected genes, we thought about using the principle of the wrapper selection approach as the next step in which we perform a Search for the Highest Accuracy on the ranked genes.

The Search for the Highest Accuracy step includes only genes which increase the accuracy. It starts with one gene and incrementally adding the other genes from the subset selected in the first step. As each gene is added in, the classifier is evaluated with the subset of genes kept.

From the first and the second step we select the most informative and useful genes for the classification task. The problem is those informative genes are not independent, and it is important to study their combined effect. The purpose of the next step is to select a minimum of informative relevant and independent genes.

The third step is Selecting a minimum subset of genes. At this stage we analyze the information provided by the constructed subset of genes. This subset is designated to be the minimum one that correctly classifies the maximum number of samples in a given training set.

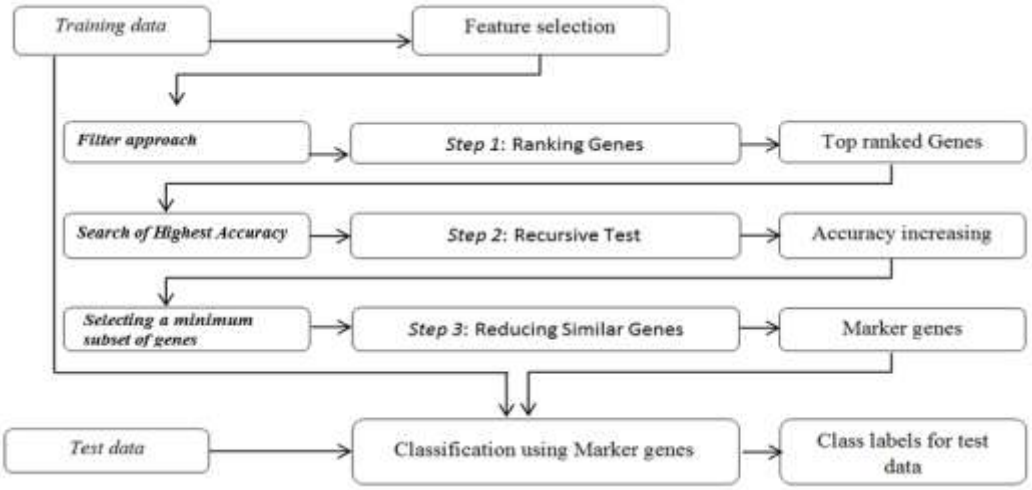

Figure 1: the new Three Filters Selection Method steps

5 A new approach for Gene Classification for binary class problems

In this section we present our new Statistic Classifier which is based on gene expression profiling measures. It is different to the existed classifiers proposed by researchers.

For Gene classification task, we propose a novel method to classify samples into binary classes.

Our new Gene classification method is based on two steps. The first step is based on the Statistic measures of the selected genes, and the second step compares the vote of genes.

For a selected gene, we search the following statistic measures in the training samples in both classes (Figure 2):

- The minimum expression value gene (Min),
- The maximum expression value (Max),
- The mean of the expression values (Mean),
- The standard deviation of the expression values (Std).

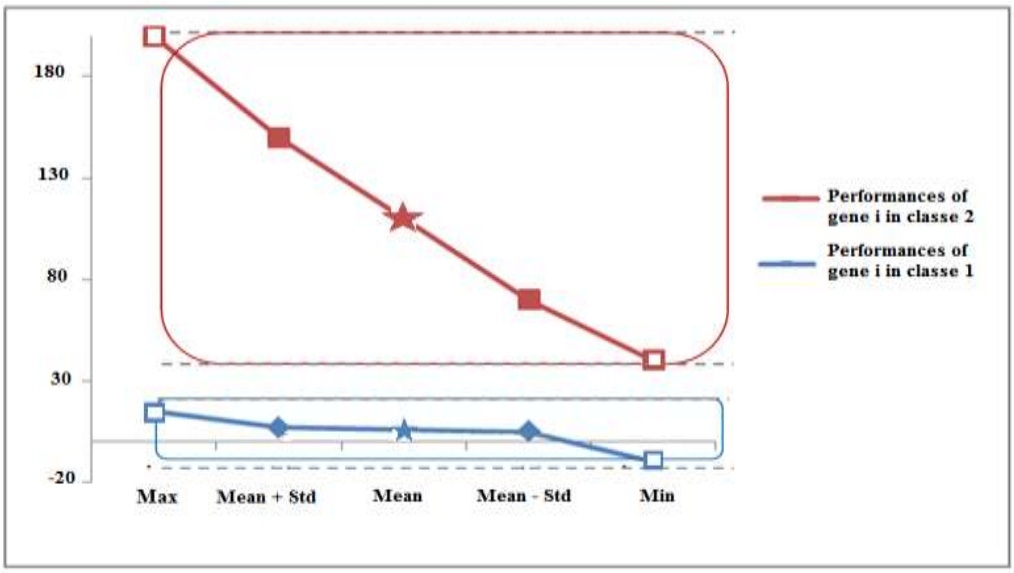

Figure 2: Map classification of The Statistic Classifier

In Figure 2, we have an example of statistic expression values for a selected gene i. The Map classification obtained helps to identify to which class belong a test sample. It contains two zones, the first one represents the zone of class 1, and the second zone belongs to the class 2.

In order to classify a test sample, we read the expression value of the gene i and try to find the right position compared to above statistic measures for both classes. The gene expression value must be between the maximum and the minimum expression values of the class (class 1 or class 2). If the gene value is in the field of class 1, the sample will be affected to the class 1, analogously for class 2.

In case where the gene expression value of a sample test is out of both zones of class 1 and class 2, then we affect the gene to the closest class especially to the interval Mean+ Std and the Mean- Std.

After this operation, each sample is classified according to selected genes.

The problem is that the selection step selects several genes, and genes may classify a test sample differently.

The suggested solution is to give a vote to each gene. The most voted class will be the right class for the test sample.

6 Datasets

6.1 Leukemia cancer

Leukemia is a cancer that usually begins in the bone marrow and results in high numbers of abnormal white blood cells. These white blood cells are not fully developed and are called blasts or leukemia cells. There are two main types of leukemia: acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML).

Leukemia's dataset is composed of 7129 genes and 72 samples, which are all acute leukemia patients, either acute lymphoblastic leukemia (ALL) or acute myeloid leukemia (AML). More

information on this dataset can be found in [12] and data can be downloaded from the website: broadinstitute.org/cgi-bin/cancer/publications/pub_paper.cgi?mode=view&paper_id=43.

6.2 Prostate cancer

Prostate cancer is a cancer that appears in one's prostate, producing the seminal fluid which nourishes and transports sperm.

Prostate cancer's dataset contains 101 samples, 52 prostate tumors and 49 non-tumor prostate samples using oligonucleotide microarrays, containing probes for approximately 12,600 genes. A more complete description of this dataset can be found in [40] and data can be downloaded from this website: broadinstitute.org/cgi-bin/cancer/publications/pub_paper.cgi?mode=view&paper_id=75.

6.3 Colon cancer

Colon cancer, known as colorectal cancer, rectal cancer or bowel cancer, is the development of cancer in the colon or rectum. It is due to the abnormal growth of cells that have the ability to invade or spread to other parts of the body.

This data set contains gene expression in 40 tumors and 22 normal colon tissues, samples were analyzed with an Affymetrix oligonucleotide array complementary to more than 6,500 human genes. Two thousand out of around 6500 genes were selected based on the confidence in the measured expression levels. A more complete description of this dataset can be found in [41] and data can be downloaded from this website: genomics-pubs.princeton.edu/oncology/affydata/insdex.html

7 Results & discussion of the new Gene Selection method

In this section we present results obtained by using Matlab to simulate the programs.

We used three datasets to study the performances of the methods used in this paper. The three datasets are Leukemia, Prostate cancer, and Colon cancer.

We divided each data set into training samples and test samples. Training samples are used to select relevant genes and to construct an adequate classification model;While, The test samples are used to test the performances of the subset of genes selected and to test the constructed classifier.

We used three filter selection methods (SNR, CC and ReliefF) to select relevant genes for each dataset, and we classified cancers by the use of three classifiers (KNN, SVM, and LDA).

For each cancer, we applied our Three Filters Selection Method (3FSM) to select the most informative and relevant genes, on the filter selection methods SNR, CC and ReliefF.

In the following tables, we present accuracies obtained by each classifier in percentages, and the corresponding number of selected genes is in parentheses.

Table 1 presents results of leukemia cancer. The adopted classifiers give accuracies between 97 and 100%. Selection methods select between 2 and 93 genes.

The SNR selection method gives accuracies between 97 and 100% for 3 to 13 genes. Our 3FSM approach increase accuracies to 100% for 3 genes.

The CC selection method gives accuracies between 97 and 100% for 3 to 93 genes. Our 3FSM approach increase accuracies to 100% for 4 genes.

The ReliefF selection method gives accuracies which reach 97% for 2 to 69 genes. Our 3FSM approach increase accuracies to 100% for only 3 genes.

Our 3FSM approach keeps only the most relevant and informative genes by excluding redundant and noisy genes. Consequently, our approach decreases the size of the selected subset of genes without any loss in the information presented by the original set of genes.

Table 1: Leukemia cancer Performances

		Classifiers		
	Selection methods	KNN	SVM	LDA
leukemia cancer	SNR	100 % (13)	97 % (4)	97 % (9)
	SNR_3FSM	100 % (3)	97 % (2)	97% (4)
	CC	100 % (50)	97% (3)	100 % (93)
	CC_3FSM	100 % (4)	97 % (3)	100 % (5)
	ReliefF	97 % (41)	97% (2)	97 % (69)
	ReliefF_3FSM	100% (4)	97 % (1)	100 % (4)

Table 2 presents results of Prostate cancer. The three filter selection methods select between 4 and 75 genes. The three classifiers give between 85 to 92%.

The SNR selection method gives accuracies between 90 and 92% for 4 to 22 genes. Our 3FSM approach increases accuracies to 95% for 3 genes.

The CC selection method gives accuracies between 85 and 92% for 6 to 44 genes. Our 3FSM approach increase accuracies to 92% for 4 genes.

The ReliefF selection method gives accuracies between 90 and 92% for 32 to 75 genes. Our 3FSM approach increase accuracies to 100% for 4 to 5 genes.

Our 3FSM approach increases The Prostate cancer classification accuracy from 85 to 95%, and decrease the number of selected genes from 75 to 3 genes.

Table 2: Prostate cancer Performances

		Classifiers		
	Selection methods	KNN	SVM	LDA
Prostate cancer	SNR	90 % (22)	92 % (8)	92 % (4)
	SNR_3FSM	95 % (3)	95 % (3)	95 % (3)
	CC	85 % (6)	92 % (44)	92 % (6)

CC_3FSM	92 % (4)	95 % (4)	95 % (4)
ReliefF	90 % (32)	92 % (34)	91 % (75)
ReliefF_3FSM	95 % (5)	95 % (4)	91 % (4)

Table 3 presents results of Colon cancer. The range of obtaining accuracies is 78.5 to 92.8% for 2 to 78 genes.

The SNR selection method gives accuracies between 85.7 and 92.8% for 2 to 29 genes. Our 3FSM approach increase accuracies to 96% for 4 to 5 genes.

The CC selection method gives accuracies between 85.7 and 92.8% for 2 to 27 genes. Our 3FSM approach increase accuracies to 96% for 5 genes.

The ReliefF selection method gives accuracies between 78.5 and 85.7% for 11 to 78 genes. Our 3FSM approach increase accuracies to 92.8% for 4 genes.

Our 3FSM approach improves the obtained accuracies for Colon cancer. It increases the classification accuracy from 78.5 to 96%, and decreases the number of selected genes from 78 to 4-5 genes.

Table 3: Colon cancer Performances

		Classifiers		
	Selection methods	KNN	SVM	LDA
Colon cancer	SNR	92.8 % (5)	85.7 % (29)	92.8 % (2)
	SNR_3FSM	96 % (4)	92.8% (4)	94 % (5)
	CC	92.8 % (7)	85.7 % (2)	92.8% (27)
	CC_3FSM	96 % (5)	95 % (3)	95 % (4)
	ReliefF	85.7 % (40)	85.7% (11)	78.5 % (78)
	ReliefF_3FSM	91 % (5)	92.8% (4)	91 % (4)

We deduce from genes selection and cancers classification that our 3FSM approach selects only relevant genes which gives the highest classification accuracies. Our adopted method improves filter selection methods and minimizes the size of the final relevant subset of genes.

As a result, our new approach selects only 3 genes for Leukemia, 3 genes for Prostate cancer, and 4 for Colon cancer.

After selecting the most significant and relevant genes for cancer classification, we remarked that the combination between the SNR filter selection method and our approach, give the highest performances for a reduced subset of relevant genes.

8 Results & discussion of the new Gene classification method

In this section we perform a comparative study between the results of our Statistic Classifier and the KNN, SVM, and LDA classifiers.

We used the selected subset by the signal to noise ratio and our 3FSM (SNR_3FSM). The same subset is used to train the four classifiers: KNN, SVM, LDA, and SC.

In Tables 4 to 6, we present obtained accuracies and the executed time for each classifier.

Table 4: The Statistic Classifier performances: Application on Leukemia

		Classifiers			
Leukemia	Selection method	KNN	SVM	LDA	SC
	SNR_3FSM	100 %	97 %	97%	100%
	Executed time	2.3s	2.4s	3.1s	1.9s

We applied different selection methods to detect the most relevant genes for Leukemia classification. The method that gives the highest performances is our approach SNR_3FSM, which selects only 3 genes.

Table 4 presents classification accuracies of Leukemia classification, for the 3 relevant genes.

The KNN classifier gives an accuracy of 100%; we obtain 97% with SVM and LDA. Our approach SC gives the highest accuracy which reaches to 100% in only 1.9s.

Table 5: The Statistic Classifier performances: Application on Prostate cancer

		Classifiers			
Prostate cancer	Selection method	KNN	SVM	LDA	SC
	SNR_3FSM	95 %	95 %	95 %	99.3%
	Executed time	2.3s	2.4s	3.1s	1.9s

In the Table 5, we present results of Prostate cancer classification. We used SNR_3FSM to select the most reduced subset of the relevant genes for Prostate cancer classification. Then we applied different classifiers to compare their performances.

For 3 genes, the KNN, SVM, and LDA classifiers gives an accuracy of 95%. Our SC classifier gives 99.3% in only 1.9s, the highest accuracy of classification for Prostate cancer.

Table 6: The Statistic Classifier performances: Application on Colon cancer

		Classifiers			
	Selection method	KNN	SVM	LDA	SC
Colon cancer	SNR_3FSM	96 %	92.8 %	94 %	97%
	Executed time	2.3s	2.5s	3.1s	1.9s

The SNR_3FSM selection method selects the most reduced subset of 4 relevant genes for Colon cancer. In the Table 6, we present the according accuracies obtained by applying three classifiers: KNN, SVM, and LDA.

The selected subset gets as accuracy 96% by the use of KNN. We obtained 94% by adopting LDA, and 92.8% by SVM. Our SC gives the highest performances which reach 97% in only 1.9s.

9 Conclusion

Two main problems exist in the process of Microarray Gene classification: Gene Selection and Gene classification. On the one hand, the Gene selection process removes redundant and irrelevant genes. And consequently, reduces the dimensionality of the dataset to be processed by the classifier and improves the predictive accuracy. On the other hand, The Gene classification uses gene expression profiling to predict the diagnosis of test samples, by generating a classify model from labeled gene expression training samples.

In this paper, we presented two approaches: the first one is the Gene Selection approach the so-called Three Filter selection method which is composed of three filters. The first filter is a direct use of a Filter selection approach. The second filter searches for the highest accuracy and removes noisy genes. And the third filter selects a minimum subset of genes and removes redundancy genes. The second one is the Gene Classifier the so-called Statistic Classifier, based on gene expression statistical measures. It classifies a test sample based on genes's statistical information on all the existed classes. Each gene proposes a class of the sample data, and the classifier decides the final class through taking the most voted class by the majority of genes.

We tested both Selection and Classification approaches on well-known cancer data sets. We used Leukemia, Prostate, and Colon cancers. Then we performed the three Filters Selection method by using the SNR, CC, or ReliefF as the first filter. Afterwards, we removed noisy genes by the use of the wrapper strategy combined with one of the classifiers: KNN, SVM, or LDA. Finally, we filtered remaining redundant genes by keeping the minimum subset of relevant genes.

The Obtained results have shown that the use of SNR as a first filter permits the selection of a minimum subset of relevant genes which gets the highest classification accuracy. After this step, we tested the performances of our Statistic Classifier by comparing obtained accuracies with the results of KNN, SVM, or LDA classifiers. We remarked that it presents the highest accuracies compared to the three known classifiers.

The advantage of our Three Filters Selection Method and our Statistic Classifier can be subsumed in the purpose of Gene Classification, which is selecting the minimum subset of relevant genes for a cancer while obtaining the highest classification performances.

References

[1]. Sara Tarek, Reda Abd Elwahab, Mahmoud Shoman. Gene expression based cancer classification. Egyptian Informatics Journal. Volume 18, Issue 3, November 2017, pages 151-159

[2]. Chandrashekar G, Sahin F. A survey on feature selection methods. Computers and Electrical Engineering. Volume 40, 2014, Pages 16-28

[3]. C. Devi ArockiaVanithaD, Devaraj, M.Venkatesulu. Gene Expression Data Classification Using Support Vector Machine and Mutual Information-based Gene Selection. Procedia Computer Science. Volume 47, 2015, Pages 13-21

[4]. Ji Gang Zhang and Hong-Wen Deng. Gene selection for classification of microarray data based on the Bayes error. BMC Bioinformatics 2007, Volume 8, Number 1, Page 1

[5]. Ramón DíazUriarte and Sara Alvarez de Andrés. Gene selection and classification of microarray data using random forest. BMC Bioinformatics. 2006, Volume 7, Number 1, Page 1

[6]. S Kar, KD Sharma, M Maitra. Gene selection from microarray gene expression data for classification of cancer subgroups employing PSO and adaptive K-nearest neighborhood technique. Expert Systems with Applications, Volume 42 Issue 1, January 2015, pages 612-627

[7]. B.H Munro (2004): Statistical Methods for Health Care Research. 2004, Lippincott Williams & Wilkins

[8]. J Apolloni, G Leguizamón, E Alba. Two hybrid wrapper-filter feature selection algorithms applied to high dimensional microarray experiments. Applied Soft Computing, Elsevier, Volume 38, January 2016, pages 922–932

[9]. Dudoit S, Laan M, Keles S, Cornec M: Unified cross-validation methodology for estimator selection and application to genomic. Bulletin of the International Statistical Institute, 54th Session Proceedings. 2003, LX (Book 2): pages 412-415

[10]. Xiaoming Liu, Jinshan Tang. Mass Classification in Mammograms Using Selected Geometry and Texture Features, and a New SVM-Based Feature Selection Method. IEEE Systems Journal. Volume: 8, Issue: 3, Sept.2014, pages 910 – 920

[11]. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES: Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science. Oct. 1999, 286: pages 531-537.

[12]. Slonim D, Tamayo P, Mesirov J, Golub T, Lander E : Class prediction and discovery using gene expression data. Proceedings of the 4th Annual International Conference on Computational Molecular Biology (RECOMB). 2000, Universal Academy Press, Tokyo, Japan, pages 263-272.

[13]. Baha Sen, Musa Peker, Abdullah Cavusoglu, Fatih V Çelebi. A Comparative Study on Classification of Sleep Stage Based on EEG Signals Using Feature Selection and Classification Algorithms. Journal of Medical Systems. March 2014, pages 38:18

[14]. Kohavi R and John G H. Wrappers for feature subset selection. Artificial Intelligence. Volume 97, Issues 1–2, December 1997, pages 273-324

[15]. S. Kwon, H. Lee, S. Lee. Image enhancement with Gaussian filtering in time domain microwave imaging system for breast cancer detection. Electronics Letters. 2016, Volume 52, Issue 5, pages 342-344

[16]. Franck Rapaport, Andrei Zinovyev, Marie Dutreix, Emmanuel Barillot, and Jean-Philippe Vert. Classification of microarray data using gene networks. BMC Bioinformatics.. Volume 8, 2007, Pages 8: 35

[17]. Hana Salem, GamalAttiya and Nawal El-Fishawy. Classification of human cancer diseases by gene expression profiles. Applied Soft Computing, Volume 50, January 2017, Pages 124-134

[18]. Marko Robnik Sikonja, Igor Kononenko. Theoretical and empirical analysis of relieff and rrelieff. Machine Learning Journal. Volume 53. 2003. Pages 23-69

[19]. Haddou Bouazza S., Auhmani K., Zeroual A., Hamdi N. (2018). Cancer Classification Using Gene Expression Profiling: Application of the Filter Approach with the Clustering Algorithm. In: Abraham A., Haqiq A., Muda A., Gandhi N. (eds). Proceedings of the Ninth International Conference on Soft Computing and Pattern Recognition (SoCPaR 2017). SoCPaR 2017. Advances in Intelligent Systems and Computing, volume 737. Springer, Cham

[20]. AM AbdelZaher, AM Eldeib. Breast cancer classification using deep belief networks. Expert Systems with Applications, Volume 46, 2016, pages 139-144

[21]. K Kira and L. Rendell. A practical approach to feature selection. ML92 Proceedings of the ninth international workshop on Machine learning. Pages 249-256, 1992.

[22]. Haddou Bouazza. S, Auhmani.K, Zeroual.A, Hamdi.N. Selecting significant marker genes from microarray data by filter approach for cancer diagnosis. Procedia Computer Science, Volume 127, 2018, Pages 300-309

[23]. Chuan Liu, Wenyong Wang, Qiang Zhao, Xiaoming Shen, Martin Konan, A new feature selection method based on a validity index of feature subset, Pattern Recognition Letters, ISSN: 0167-8655, Volume 92,2017, Page: 1-8

[24]. Marczyk M, Jaksik R, Polanski A, Polanska J. Adaptive filtering of microarray gene expression data based on gaussian mixture decomposition. BMC Bioinformatics. Volume 14, 2013, Number 1, Page 1

[25]. M. Dashtban, Mohammadali Balafar. Gene selection for microarray cancer classification using a new evolutionary method employing artificial intelligence concepts. Genomics. Volume 109, 2017, pages 91–107

[26]. M.Dashtban, MohammadaliBalafar, PrashanthSuravajhala. Gene selection for tumor classification using a novel bio inspired multi-objective approach. Genomics. Volume 110, Issue 1, January 2018, Pages 10-17

[27]. Haddou Bouazza. S, Hamdi.N, Auhmani.K, Zeroual.A. Gene-expression-based cancer classification through feature selection with KNN and SVM classifiers. Intelligent Systems and Computer Vision (ISCV), 2015

[28]. Osama Mahmoud, Andrew Harrison, Aris Perperoglou, Asma Gul, Zardad Khan, Metodi V Metodiev, and Berthold Lausen. A feature selection method for classification within functional genomics experiments based on the proportional overlapping score. BMC Bioinformatics. Volume 15, 2014; 15(1): 274

[29]. Lausen B, Hothorn T, Bretz F, Schumacher M. Assessment of optimal selected prognostic factors. Biom J. Volume 46, 2004, Issue 3, Pages 364–374.

[30]. Haddou Bouazza. S, Auhmani.K, Zeroual. Gene expression data analyses for supervised prostate cancer classification based on feature subset selection combined with different classifiers. 5th International Conference on Multimedia Computing and Systems (ICMCS), 2016

[31]. Ultsch A, Pallasch C, Bergmann E, Christiansen H. A comparison of algorithms to find differentially expressed genes in microarray data. In: Fink A, Lausen B, Seidel W, Ultsch A, editors. Advances in Data Analysis, Data Handling and Business Intelligence. Studies in Classification, Data Analysis, and Knowledge Organization. Berlin Heidelberg: Springer; 2010. pages. 685–697.

[32]. Lu J, Kerns RT, Peddada SD, Bushel PR. Principal component analysis-based filtering improves detection for affymetrix gene expression arrays. Nucleic Acids Res. Volume 39, 2011, Issue 13, Pages 86–86.

[33]. M. Xiong, X. Fang, J. Zhao, Biomarker identification by feature wrappers, Genome. Res. Volume 11, 2001, Pages 1878-1887

[34]. G. Chen, J. Chen, A novel wrapper method for feature selection and its applications, Neurocomputing, Volume 159, 2015, pages 219-226

[35]. S. Huda, M. Abdollahian, M. Mammadov, J. Yearwood, S. Ahmed, I. Sultan, A hybrid wrapper filter approach to detect the source(s) of out-of-control signals in multivariate manufacturing process, Eur. J. Oper. Res. Volume 237, 2014, Pages 857–870

[36]. J.M. Cadenas, M.C. Garrido, R. Martinez, Feature subset selection filter wrapper based on low quality data, Expert Syst. Appl. Volume 40, 2013, Pages 621–625

[37]. Li-Yeh C, Chao-Hsuan K, Cheng-Hong Y: A Hybrid Both Filter and Wrapper Feature Selection Method for Microarray Classification. CORR, 2016

[38]. Baralis E, Bruno G, Fiori A. Minimum number of genes for microarray feature selection. Engineering in Medicine and Biology Society. EMBS 2008. 30th Annual International Conference of the IEEE. Vancouver: IEEE, 2008, Pages 5692–5695

[39]. Apiletti D, Baralis E, Bruno G, Fiori A. Maskedpainter: feature selection for microarray data analysis. Intell Data Anal. Volume 16, 2012, Issue 4, Pages 717–737

[40]. Dinesh Singh, Phillip G. Febbo, Kenneth Ross, Donald G. Jackson, Judith Manola, Chris-tine Ladd, Pablo Tamayo, Andrew A. Renshaw, Anthony V. D'Amico, Jerome P. Richie, Eric S. Lander, Massimo Loda, Philip W. Kantoff, Todd R. Golub, William R. Sellers. Cancer Cell: Volume 1, March 2002

[41]. Chanho Park, Sung Bae Cho. Evolutionary ensemble classifier for lymphoma and colon cancer classification. Conference: Evolutionary Computation, 2003, CEC'03