International Journal of Computer Information Systems and Industrial Management Applications.
ISSN 2150-7988 Volume 10 (2018) pp. 068-086

Received: 19 Dec, 2017; Accept 23 Feb, 2018; Publish: 19 April, 2018

Application of the Filter approach and the Clustering algorithm on Cancer datasets

SARA HADDOU BOUAZZA[1], KHALID AUHMANI[2], ABDELOUHAB ZEROUAL[1]

[1]Department of Physics, Faculty of Sciences Semlalia, Cadi Ayyad University, Marrakech, Morocco
Sara.hb.sara@gmail.com, zeroual@uca.ma

[2]Department of Industrial Engineering, National School of Applied Sciences, Cadi Ayyad, Safi, Morocco
kauhmani@yahoo.com

***Abstract*: In this paper, we compare the accuracy of classification for different cancers, based on gene microarray expression data. For this reason, we have used a combination between filter selection methods and clustering algorithms to select relevant features, in each cancer dataset, for gene classification.**

Our effort is carried out in two steps. First, we survey the effect of the selection methods, on the classification accuracy for cancers, by comparing the performances evaluated by different classifiers. The considered selection methods in this paper are SNR, ReliefF, Correlation Coefficient, Mutual Information, T-Statistics, Fisher, Max relevance Min redundancy, and Principal component analysis. We evaluated the performances of each selection method by the use of the K Nearest Neighbor, Support Vector Machine, Linear Discriminant Analyses, Decision tree for classification and Naïve Bayes classifier for a supervised classification task.

As a second step, we preceded the selection step by a k-means and k-medians clustering operation.

Obtained accuracies detect that the best classification accuracies were reached for a minimum subset of selected genes, in all cancers, in case we applied the k-means clustering for the selected genes by the filter methods.

***Keywords*:** DNA Microarray; Feature selection; Supervised Classification; Clustering; image processing; Cancer classification.

I. Background

DNA microarrays are characterized the high number of genes and a limited number of samples. For this reason, it is necessary to reduce the dimensionality of dataset to make the classification task clearer, easier and faster.

The most common form for dimensionality reduction is feature subset selection, an imperative process for cancer classification.

To classify a cancer dataset, we most select relevant features which best represent the cancer dataset.

In this paper, we suggest to use the k means clustering as a selection method. We combined between filter selection methods and clustering algorithms. To compare these feature selection methods, an evaluation of the dimensionality reduction had been done using seven supervised classifiers

The goal of this combination is to improve classification performance and to accelerate the search to identify relevant feature subsets.

II. Related Works

Features selection methods become the focus of much research in areas of application for which datasets with thousands of features are available. Some of the used methods in the field of feature selection are:

- The use of the random forest (RF) which constructs multiple decision tree [1].
- The proposed method improves the stability of the wrapper variable selection procedures while preserves and possibly improves the classification performance [2].
- The use of the feature selection technique of Filter-Embedded Feature Ranking Techniques (FEFR), which is the combination of the filter method (ReliefF) and embedded methods (Variable Importance based Random Forest) by [3].
- Fisher, T-statistics, Signal to noise ratio and ReliefF selection methods [4].
- The use of two-step neural network classifier [5].
- The (BW) discriminant score was proposed by [6]. It is based on the dispersion ratio between classes and intra-class dispersion.
- A hybridization between Genetic Algorithm (GA) and Max-relevance, Min-Redundancy (MRMR) [7]

III. Materials and Methods

To prove the importance of the k-means clustering step, we used different feature selection methods and classifiers for cancer classification.

In the first step, we used dataset of different cancers composed of thousands of features. In the second step we reduced the number of features, using a feature subset selection, to only relevant features. In the final step, we classify the datasets.

A. Dataset Description

In this paper, we investigate the effect of feature selection methods on six commonly used gene expression datasets: leukemia cancer, Colon cancer and Prostate cancer, Lung cancer, Lymphoma cancer, and CNS cancer (table 1).

- Leukemia is composed of 7129 genes and 72 samples. It contains two classes: acute lymphocytic leukemia (ALL) and acute myelogenous leukemia (AML). It can be downloaded from the website[1]
- Colon cancer is composed of 6500 genes and 62 samples. It contains two classes: Tumor and Not tumor. It can be downloaded from this website[2]
- Prostate cancer is composed of 12600 genes and 101 samples. It contains two classes: Tumor and Not tumor. It can be downloaded from this website[3]
- Lung Cancer is composed of 12533 genes and 181 samples; it contains two classes: malignant pleural mesothelioma (MPM) and adenocarcinoma (ADCA). Data could be downloaded from the website[4]
- Lymphoma cancer is composed of 7070 genes and 77 samples. It contains two classes: diffuse large B-cell lymphoma (DLBCL) and follicular lymphoma (FL). It is available to the public at the website[5]
- The central nervous system (CNS) is the part of the nervous system consisting of the brain and spinal cord. The CNS Tumor dataset contains information about 60 patients, 21 patients died and 39 survived, for each experiment we have 7129 gene expression values. For more information about these data you can visit the website[6]

Dataset	No. of features	No. of observation	No. of classes
Leukemia [8]	7129	72	2
Colon [9]	6500	62	2
Prostate [10]	12600	101	2
Lung [11]	12533	181	2
Lymphoma [12]	7070	77	2
Central nervous system [13]	7129	60	2

Table 1. Datasets and parameters used for experiments

B. Feature Subset Selection

Feature selection is the operation of selecting relevant genes for cancer classification [14] (figure1).

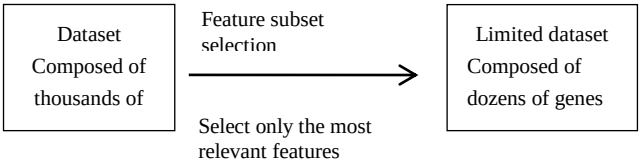

Figure 1. Feature subset selection

The feature selection process is the task of selecting relevance genes by removing It exists three main categories of feature selection algorithms: wrappers, filters and embedded methods [15].

- Wrapper methods use a predictive model to score feature subsets.
- Filter methods use a proxy measure instead of the error rate to score a feature subset.
- Embedded methods are a catchall group of techniques which perform feature selection as part of the model construction process.

In this paper, we used filter methods which are based on the estimated weight for each gene, to select the relevant subset of genes for cancer classification.
The methods used in this work are the SNR, ReliefF, Correlation Coefficient, Mutual Information, T-Statistics, Fisher, Max relevance Min redundancy, Principal component analysis, and clustering k-means and k-medians.

1) The signal to noise ratio

The signal to noise ratio, calculate the score S/R of each gene (g) [16] [8] as follows:

$$S/R_{(g)} = \frac{M_{1g} - M_{2g}}{S_{1g} + S_{2g}} \quad (1)$$

Where M_{kg} and S_{kg} denote the mean and the standard deviation of the feature g for samples of classes 1 and 2

2) ReliefF

This algorithm presented as Relief [17] and adjusted to the multi-class case by Kononenko as the ReliefF [18].
This criterion measures the ability of each feature to group data of the same class and discriminating those having different classes. The algorithm is described as follows:

- Initialize the score (or the Weight) wd=0, d=1, .., D
- For t = 1 …N
- Pick randomly an instance x_i
- Find the k nearest neighbors to x_i having the same class (hits)
- Find the k nearest neighbors to x_i having different class (misses c)
- For each feature d, update the weight:

$$Wd = wd - \sum_{j=1}^{K} \frac{diff(x_i,d,hits_j)}{m*k} + \sum_{c \neq class(x_i)} \frac{p(c)}{1-p(class(x_i))} \sum_{j=1}^{k} \frac{diff(x_i,d,misses_j)}{m*k} \quad (2)$$

The distance used is defined by:

[1] broadinstitute.org/cgi-bin/cancer/publications/pub_paper.cgi?mode=view&paper_id=43
[2] genomics-pubs.princeton.edu/oncology/affydata/insdex.html
[3] broadinstitute.org/cgi-bin/cancer/publications/pub_paper.cgi?mode=view&paper_id=75
[4] http://www.chestsurg.org
[5] http://www.broadinstitute.org/cgi-bin/cancer/datasets.cgi
[6] http://csse.szu.edu.cn/staff/zhuzx/Datasets.html

$$\text{diff}(x_i, d, x_j) = \frac{|x_{id} - x_{jd}|}{\max(d) - \min(d)} \quad (3)$$

Max (d) (resp. min (d)) is the maximum (resp. minimum) value that may take the feature designated by the index d on the data set. xid is the value of the dth feature of the data xi.
This method does not eliminate redundancy, but defines a relevant criterion.

3) T Statistics

The calculated score "t" for each feature (g) is used in [19]:

$$t_{(g)} = \frac{X_1 - X_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \quad (4)$$

Where nk, Xk and Sk^2 are the size, the average and the variance of classes k = 1, 2.

4) F test

The F test gives a score defined as follows [20]:

$$F(g) = \frac{(M1 - M2)^2}{(S1^2 + S2^2)} \quad (5)$$

Where Mk; Sk^2 denotes the mean and standard deviation of the feature (g) for the class k = 1; 2.

5) Correlation Coefficient.

Correlation coefficients measure the strength of association between two features [21].

Let S and S_y be the standard deviations of two random features X and Y respectively. Then the Pearson's product moment correlation coefficient between the features is:

$$\rho_{x,y} = \frac{cov(X,Y)}{S_x S_y} = \frac{E((X - E(X))(Y - E(Y)))}{S_x S_y} \quad (6)$$

Where cov(.) means covariance and E(.) denotes the expected value of the feature.

6) Max-relevance, Min-Redundancy

Minimum redundancy feature selection is an algorithm frequently used in a method to identify characteristics of genes and phenotypes and narrow down their relevance and is usually described in its pairing with relevant feature selection as Minimum Redundancy Maximum Relevance (mRMR).
Let U ={X1, X2...} denote a set of one-dimensional discrete random variables, C= {c1, c2,...,ck} is a distinguished class variable, and S U represent any subset of U.
The first principle of mRMR is that we should not use features which are highly correlated among themselves [22]; the redundancy between features should be taken into account, thus keeping features which are maximally dissimilar to each other. A way of globally measuring redundancy among the variables in S is:

$$WI(S) = 1/|S|^2 \sum_{Xi, Xj \epsilon S} MI(Xi, Xj) \quad (7)$$

Where (Xi Xj) is the measure of mutual information between the variables Xi and Xj.
The second idea of mRMR is that minimum redundancy should be supplemented by the use of a maximum relevance criterion of the features with respect to the class variable. A measure of global relevance of the variables in S with respect to C is:

$$VI(S) = 1/|S| \sum_{Xi \epsilon S} (C, Xj) \quad (8)$$

To combine redundancy and relevance we use:

$$S^* = \arg\max S \subseteq U (VI(S) - WI(S)) \quad (9)$$

The selected subset is obtained in an incremental way, starting with the feature having a maximum value of (C; Xi) (S0 = {Xi0}) and progressively adding to the current subset Sm-1 the feature which maximizes: max X j ϵ U/Sm-1 MI(C, X j) - 1/(m-1)∑ XiϵSm-1 MI(X j ,Xi)).

7) Mutual Information.

Let us consider a random feature G that can take n values over several measures, we can empirically estimate the probabilities P(G1), ..., P(Gn) for each state G1, ..., Gn of feature G. Shannon's entropy [23] of the feature is defined as:

$$H_{(G)} = -\sum_{i=0}^{NG} P_{(G)} \log (P_{G(i)}) \quad (10)$$

The mutual information measures the dependence between two features. In the situation of genes selection, we use this measure to recognize genes which are related to the class C. The mutual information between C and one gene G is measured by the following expression:

$$MI(G,C) = H(G) + H(C) - H(G,C) \quad (11)$$

$$H(G, C) = -\sum_{i=0}^{NG} -\sum_{j=0}^{NG} P_w(i, j) \log (P_w(i, j)) \quad (12)$$

8) k-means.

In clustering, Cluster analysis is the task of regrouping similar objects in groups [24]. The k-means algorithm is used to divide the samples into k groups called clusters and returns the index of the cluster to which it has assigned each feature [25]. Cancer classification using gene expression profiling: application of the filter approach with the clustering algorithm. K-means algorithm is described as two steps [26]:

- Assignment step: Assign each feature to the cluster whose mean yields the least within-cluster sum of squares.
- Update step: Calculate the new means to be the centroids of the features in the new clusters.

9) K medians

The K-medians clustering [4] [5] is a cluster analysis algorithm. It is a variation of k-means clustering where instead of calculating the mean for each cluster to determine its centroid, one instead calculates the median [27].

C. Classification

The DNA Microarray technology has proven to be encouraging in predicting cancer classification and prognosis outcomes [28]. The DNA Microarray classification uses gene expression array phenotype to predict the diagnosis of a sample. It generates a classify model, from labeled gene expression data samples, to classify new data samples into different predefined diseases.
In this section, we present different classifiers used to evaluate

the dimensionality reduction done by selection methods on cancers datasets.

1) K Nearest Neighbors.

K nearest neighbors' is a classifier that stores training samples and classifies the test samples based on a similarity measure.
In K Nearest Neighbors, we try to find the most similar K number of samples as nearest neighbors in a given sample, and predict class of the sample according to the information of the selected neighbors.
We can compute the Euclidean distance between two samples by using a distance function DE(X, Y), where X, Y are samples composed of N features, such that X = {X1, …, XN }, Y = {Y1, …, YN }.

$$D_E(X, Y) = \sum_{j=1}^{k} \sqrt{(X_i^2 - Y_i^2)} \quad (13)$$

2) Support Vector Machines +9(SVM).

Support vector machines are supervised learning models used for supervised classification [29]. Support Vector Machines are based on two key concepts: the notion of maximum margin and the concept of kernel functions.

3) Linear Discriminant Analysis (LDA).

Linear Discriminant Analysis is an algorithm used in machine learning to search and find a linear combination of features that characterizes or separates two or more classes of objects [30].

4) Decision Tree for Classification (DTC)

Decision tree classifier uses a decision tree as a predictive model which predicts the class of a target sample by learning simple decision rules inferred from the data genes. It is one of the predictive modeling approaches used in data mining and machine learning [31]

5) Naïve Bayes (NB)

The Naive Bayes is a classifier that uses Bayes theorem and assume all attributes to be independent given the value of the class variable [32].
To evaluate the performances of the classifiers, we measure the value of the classification accuracy Accuracy [33]:

$$A_{ccuracy} = 100* (TP + TN) / (TN + TP+ FN+ FP) \quad (14)$$

Where TP is the 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 the false negative for incorrect prediction to normal class.
All the algorithms used in this paper have been run using (MATLAB).

IV. Results

In this section, we report results obtained from an experimental study of the effect of the k-means clustering on six commonly used gene expression datasets. Each dataset is characterized by a group of genes.
After dividing the initial dataset into training and test sample, we applied a subset selection method on training samples to select relevant genes. Then we classify dataset using the classifiers (KNN, SVM, LDA, DTC and NB). Test samples are used to investigate the performances of subset selected by selection methods (SNR, ReliefF, CC, MI, T-S, Fisher, MRmr and PCA)
To increase the selection methods performances, we add a clusterisation task to the selection step. We divide training samples into clusters, then we select relevant features in each cluster. The obtained subset presents the most relevant features in the dataset.
Tables and figures 2 to 7 compares the classification accuracy obtained for the number of genes selected (in italic) (for leukemia, colon, prostate, lung, lymphoma and CNS cancers, respectively) before and after adding the k-means and k-medians clustering to the selection step.
We can clearly remark the advantage of adding the clusterisation step to the feature selection process. It increases the accuracy of the selection methods investigated and decrease the dimensionality of the datasets.

V. Discussion

Tables and figures 2 to 7 presents accuracies obtained for the selected genes by the selection methods SNR, ReliefF, CC, MI, T-S, Fisher, MRmr, and PCA. It presents also results after adding a second selection step which is k-means and k-medians clustering.
For Leukemia cancer, we remark that the obtained results are between 100% and 44.11%. The average of accuracies is 89.92% for a number of genes between 2 and 95.
After adding the k-means to the selection step we obtain accuracies between 100% and 91.1%. The average of accuracies is 96.44% for 2 to 35 genes.
After adding the k-medians to the selection step we obtain accuracies between 100% and 91.1%. The average of accuracies is 95.92% for 1 to 42 genes.
From these results we can deduce that the k-means clustering increase accuracies with 6.52%. The k-medians increase accuracies with 6%.
For Colon cancer, accuracies are in the range of 92.8% and 71.4%. The average of accuracies is 83.89% for a number of genes between 2 and 43.
After adding the k-means to the selection step we obtain accuracies between 100% and 85.7%. The average of accuracies is 91.66% for 2 to 28 genes.
After adding the k-medians to the selection step we obtain accuracies between 100% and 78.65%. The average of accuracies is 89.44% for 2 to 28 genes.
From these results we can deduce that the k-means clustering increase accuracies with 7.77%. The k-medians increase accuracies with 5.55%.
For Prostate cancer, we remark that the obtained accuracies are between 100% and 54.9%. The average of accuracies is 79.06% for a number of genes between 1 and 75.
After adding the k-means to the selection step we obtain accuracies between 100% and 65%. The average of accuracies is 84.38% for 1 to 43 genes.
After adding the k-medians to the selection step we obtain accuracies between 100% and 58.8%. The average of accuracies is 80.88% for 2 to 52 genes.

From these results we can deduce that the k-means clustering increase accuracies with 5.32%. The k-medians increase accuracies with 1.82%.

For Lung cancer, we remark that the obtained results are between 100% and 66.4%. The average of accuracies is 93.75% for a number of genes between 1 and 82.

After adding the k-means to the selection step we obtain accuracies between 100% and 83.2%. The average of accuracies is 96.65% for 2 to 28 genes.

After adding the k-medians to the selection step we obtain accuracies between 100% and 67.7%. The average of accuracies is 95.68% for 2 to 34 genes.

From these results we can deduce that the k-means clustering increase accuracies with 2.9%. The k-medians increase accuracies with 1.93%.

For Lymphoma cancer, we remark that the obtained results are between 100% and 52.1%. The average of accuracies is 92.47% for a number of genes between 1 and 97.

After adding the k-means to the selection step we obtain accuracies between 100% and 86.9%. The average of accuracies is 95.85% for 1 to 38 genes.

After adding the k-medians to the selection step we obtain accuracies between 100% and 82.6%. The average of accuracies is 95.19% for 2 to 52 genes.

From these results we can deduce that the k-means clustering increase accuracies with 3.38%. The k-medians increase accuracies with 2.72%.

For CNS cancer, we remark that obtained accuracies are between 76.7% and 44.1%. The average of accuracies is 63.65% for a number of genes between 1 and 98.

After adding the k-means to the selection step we obtain accuracies between 84% and 58.1%. The average of accuracies is 70.19% for 2 to 35 genes.

After adding the k-medians to the selection step we obtain accuracies between 184% and 55.8%. The average of accuracies is 66.61% for 2 to 35 genes.

From these results we can deduce that the k-means clustering increase accuracies with 6.54%. The k-medians increase accuracies with 2.96%.

VI. Conclusion

We have presented in this paper that feature selection methods can be practiced successfully to the cancer classification, using simply a limited number of training samples in a high dimensional space of thousands of genes.

We performed quite a few studies on leukemia, colon, prostate, lung, lymphoma and CNS cancer datasets. The objective was to classify each cancer dataset into two classes.

The obtained results show that the proposed clustering algorithm has efficient searching strategies and is capable of selecting an important subset of genes for cancer classification while increasing accuracies and decreasing the selected subset of genes simultaneously.

For all cancers, we remarked that both k-means and k-medians do increase classification accuracies and decrease the number of selected genes. The k-means present the best improvement done for the studied filter selection methods, and also, reduces the high dimensionality of data to the most limited subset of relevant genes.

Leukemia cancer accuracies were increased by 6.52%. Colon cancer accuracies were increased by 7.77%. Prostate cancer by 5.32%. Lung cancer by 2.9%. Lymphoma cancer by 3.38%. And CNS cancer by 6.54%.

These results encourage adding a clusterisation before the selection step, and specially the k-means clustering. It increases the classification accuracies and decreases the number of features selected.

	KNN		SVM		LDA		DTC		NB	
	Acc (%)	*Nbr genes*	Acc (%)	*Nbr genes*	Acc (%)	*Nbr genes*	Acc (%)	*Nbr genes*	Acc (%)	*Nbr Genes*
SNR	100	*13*	97.05	*4*	100	*9*	97.05	*3*	97.05	*5*
ReliefF	100	*41*	97.05	*2*	100	*69*	94.11	*11*	44.11	*5*
CC	100	*50*	97.05	*2*	100	*93*	97.05	*4*	97.05	*6*
MI	76.41	*56*	84.2	*5*	91.1	*10*	76.4	*86*	91.1	*28*
T-S	97.05	*75*	97.05	*2*	97.05	*66*	91.17	*13*	58.82	*95*
Fisher	97.05	*69*	84.2	*59*	97.05	*93*	58.82	*8*	73.52	*2*
MRmr	97.05	*11*	88.2	*30*	85.2	*40*	64.7	*33*	88.2	*12*
PCA	100	*15*	97.05	*7*	100	*13*	97.05	*15*	91.1	*25*
K-means + SNR	100	*5*	100	*4*	100	*5*	97.05	*2*	97.05	*3*
K-means + ReliefF	100	*8*	100	*3*	100	*21*	97.05	*6*	91.1	*12*
K-means + CC	100	*19*	100	*12*	100	*35*	100	*3*	97.05	*5*
K-means + MI	91.1	*18*	91.1	*5*	94.1	*5*	94.1	*28*	94.1	*15*
K-means + T-S	100	*12*	100	*11*	97.05	*12*	97.05	*13*	91.1	*35*
K-means + Fisher	97.05	*6*	97.05	*5*	97.05	*13*	91.1	*6*	91.1	*12*
K-means + MRmr	97.05	*5*	91.1	*16*	94.1	*20*	91.1	*23*	91.1	*8*
K-means + PCA	100	*9*	97.05	*5*	100	*10*	100	*11*	94.1	*7*
K-medians + SNR	100	*7*	100	*5*	100	*9*	100	*5*	97.05	*4*
K-medians + ReliefF	100	*12*	97.05	*1*	100	*23*	97.05	*10*	91.1	15
K-medians + CC	100	*23*	100	*14*	100	*40*	100	*15*	97.05	5
K-medians + MI	91.1	*26*	91.1	*20*	94.1	*15*	91.1	*12*	94.1	26
K-medians + T-S	97.05	*21*	100	*15*	97.05	*16*	94.1	*24*	91.	*42*
K-medians + Fisher	97.1	*11*	97.05	*6*	97.05	*21*	91.1	*16*	91.1	*22*
K-medians + MRmr	97.05	*10*	91.1	*20*	91.1	*3*	91.1	*31*	91.1	*15*
K-medians + PCA	97.05	*9*	97.05	*5*	100	*11*	97.05	*3*	91.1	*2*

Table 2. Performance of comparison for proposed classifiers (leukemia cancer)

Figure 2. Performance of comparison for proposed classifiers (leukemia cancer)

	KNN		SVM		LDA		DTC		NB	
	Acc (%)	*Nbr genes*	Acc (%)	*Nbr genes*	Acc (%)	*Nbr genes*	Acc (%)	*Nbr genes*	Acc (%)	*Nbr Genes*
SNR	92.8	*5*	85.7	*29*	92.8	*2*	85.7	*12*	92.8	*12*
ReliefF	85.7	*40*	85.7	*11*	78.5	*7*	85.7	*11*	85.7	*18*
CC	92.8	*7*	85.7	*2*	92.8	*27*	92.8	*15*	92.8	*21*
MI	85.7	*43*	78.5	*5*	71.4	*19*	71.4	*13*	71.4	*13*
T-S	92.8	*17*	85.7	*12*	85.7	*29*	85.7	*14*	85.7	*22*
Fisher	85.7	*4*	85.7	*15*	78.5	*7*	78.5	*2*	71.4	*26*
MRmr	85.7	*3*	71.4	*5*	71.4	*19*	71.4	*13*	71.4	*13*
PCA	92.8	*16*	85.7	*2*	85.7	*21*	85.7	*7*	92.8	*10*
K-means + SNR	**95**	*6*	**100**	*4*	**100**	*8*	92.8	*2*	92.8	*2*
K-means + ReliefF	**95**	*25*	**92.8**	*7*	**92.8**	*15*	92.8	*28*	92.8	*20*
K-means + CC	**94.2**	*2*	**95**	*2*	**95**	*14*	92.8	*10*	100	*21*
K-means + MI	**95**	*25*	**91.1**	*5*	**94.1**	*3*	85.7	*3*	85.7	*23*
K-means + T-S	92.8	*7*	91.1	*12*	91.1	*2*	91.1	*4*	85.7	*2*
K-means + Fisher	91.1	*14*	91.1	*21*	85.7	*11*	85.7	*10*	85.7	*12*
K-means + MRmr	91.1	*11*	85.7	*10*	85.7	*3*	85.7	*7*	85.7	*12*
K-means + PCA	100	*13*	91.1	*12*	91.1	*2*	91.1	*17*	92.8	*3*
K-medians + SNR	92.8	*2*	91.1	*12*	100	*21*	91.1	*21*	92.8	*5*
K-medians + ReliefF	91.1	*12*	91.1	*10*	91.1	*10*	92.8	*12*	92.8	*25*
K-medians + CC	94.2	*14*	92.8	*28*	94.1	*14*	92.8	*11*	95	*14*
K-medians + MI	92.8	*15*	85.7	*17*	85.7	*12*	85.7	*14*	85.7	*25*
K-medians + T-S	92.8	*11*	85.7	*3*	91.1	*14*	91.1	*12*	85.7	*15*
K-medians + Fisher	91.1	*15*	91.1	*24*	85.7	*22*	85.7	*14*	78.5	*12*
K-medians + MRmr	85.7	*2*	78.5	*14*	78.5	*22*	78.5	*10*	85.7	*14*
K-medians + PCA	95	*12*	91.1	*21*	91.1	*12*	91.1	*11*	92.8	*11*

Table 3. Performance of comparison for proposed classifiers (colon cancer)

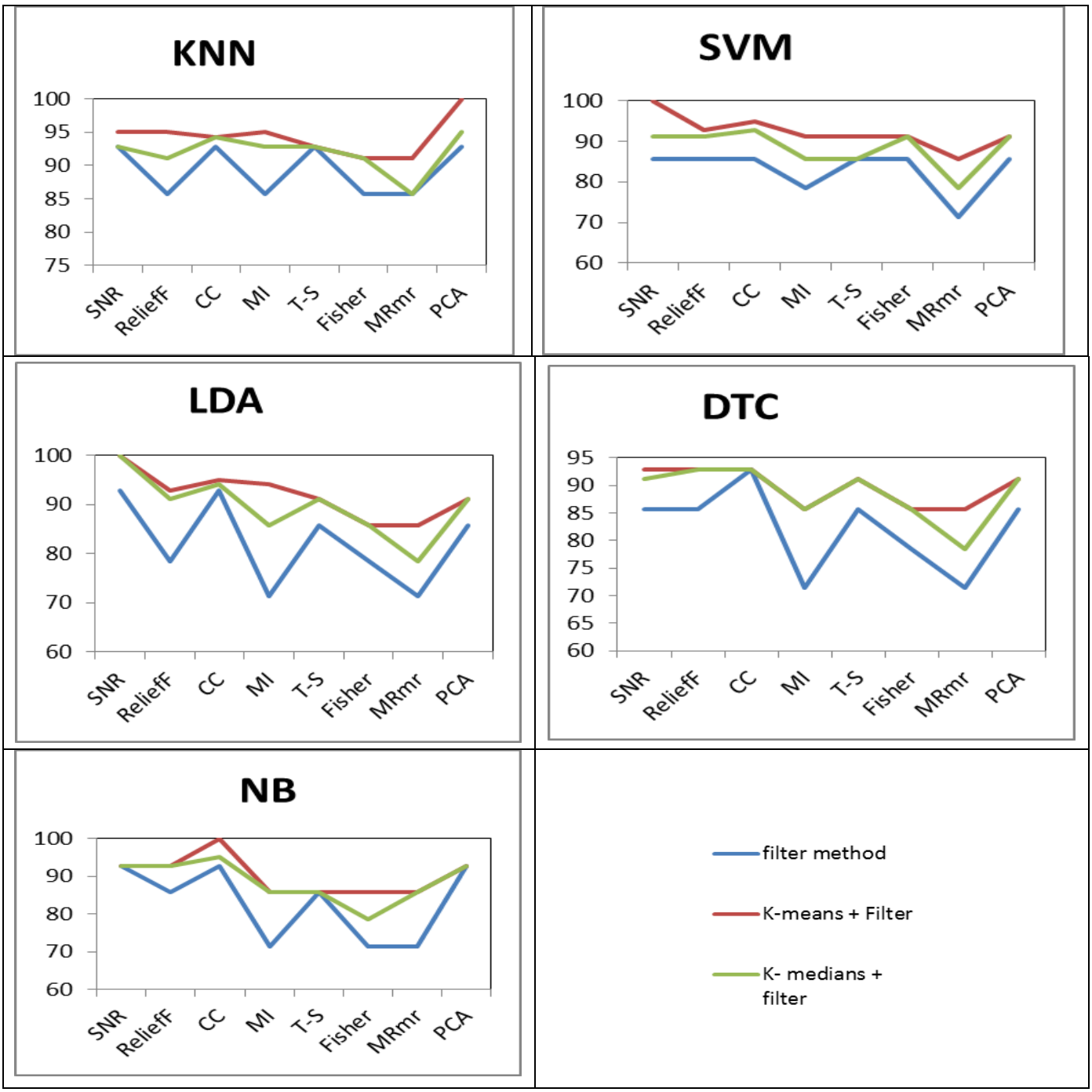

Figure 3. Performance of comparison for proposed classifiers (Colon cancer)

	KNN		SVM		LDA		DTC		NB	
	Acc (%)	*Nbr genes*	Acc (%)	*Nbr genes*	Acc (%)		Acc (%)	*Nbr genes*	Acc (%)	*Nbr genes*
SNR	90	*22*	92	*8*	100	*4*	90	*18*	90	*22*
ReliefF	90	*32*	92	*34*	100	*75*	90	*42*	90	*27*
CC	85	*6*	92	*44*	100	*6*	85	*31*	90	*37*
MI	65	*1*	58.8	*56*	92	*10*	58.8	*12*	58.8	*21*
T-S	90	*12*	78.4	*31*	78.4	*15*	65	*31*	65	*22*
Fisher	78.4	*22*	65	*12*	65	*22*	58.8	*34*	58.8	*12*
MRmr	60	*7*	68.6	*60*	65	*49*	54.9	*3*	60	*23*
PCA	92	*25*	90	*18*	90	*27*	85	*22*	85	*15*
K-means + SNR	**90**	*1*	**100**	*9*	**100**	*3*	90	*3*	90	*2*
K-means + ReliefF	**90**	*5*	**92**	*7*	**100**	*43*	90	*11*	90	*6*
K-means + CC	**90**	*1*	**92**	*5*	**100**	*3*	90	*12*	90	*10*
K-means + MI	**90**	*4*	**78.4**	*10*	**95**	*8*	65	*3*	65	*14*
K-means + T-S	92	*13*	90	*18*	90	*12*	78.4	*21*	78.4	*12*
K-means + Fisher	85	*13*	65	*3*	65	*2*	78.4	*12*	65	*13*
K-means + MRmr	65	*13*	78.4	*13*	78.4	*14*	65	*22*	65	*18*
K-means + PCA	92	*10*	92	*17*	90	*13*	85	*12*	90	*11*
K-medians + SNR	90	*12*	92	*3*	100	*3*	90	*11*	90	*12*
K-medians + ReliefF	90	*13*	92	*14*	100	*52*	90	*13*	90	*16*
K-medians + CC	85	*2*	92	*14*	100	*5*	90	*18*	90	*12*
K-medians + MI	78.4	*10*	65	*12*	92	*3*	58.8	*2*	65	*16*
K-medians + T-S	90	*3*	85	*13*	85	*10*	65	*3*	65	*2*
K-medians + Fisher	78.4	*3*	65	*7*	65	*5*	65	*12*	58.8	*3*
K-medians + MRmr	60	*3*	68.6	*10*	78.4	*16*	58.8	*12*	65	*21*
K-medians + PCA	92	*12*	90	*3*	90	*17*	85	*13*	85	*3*

Table 4. Performance of comparison for proposed classifiers (prostate cancer)

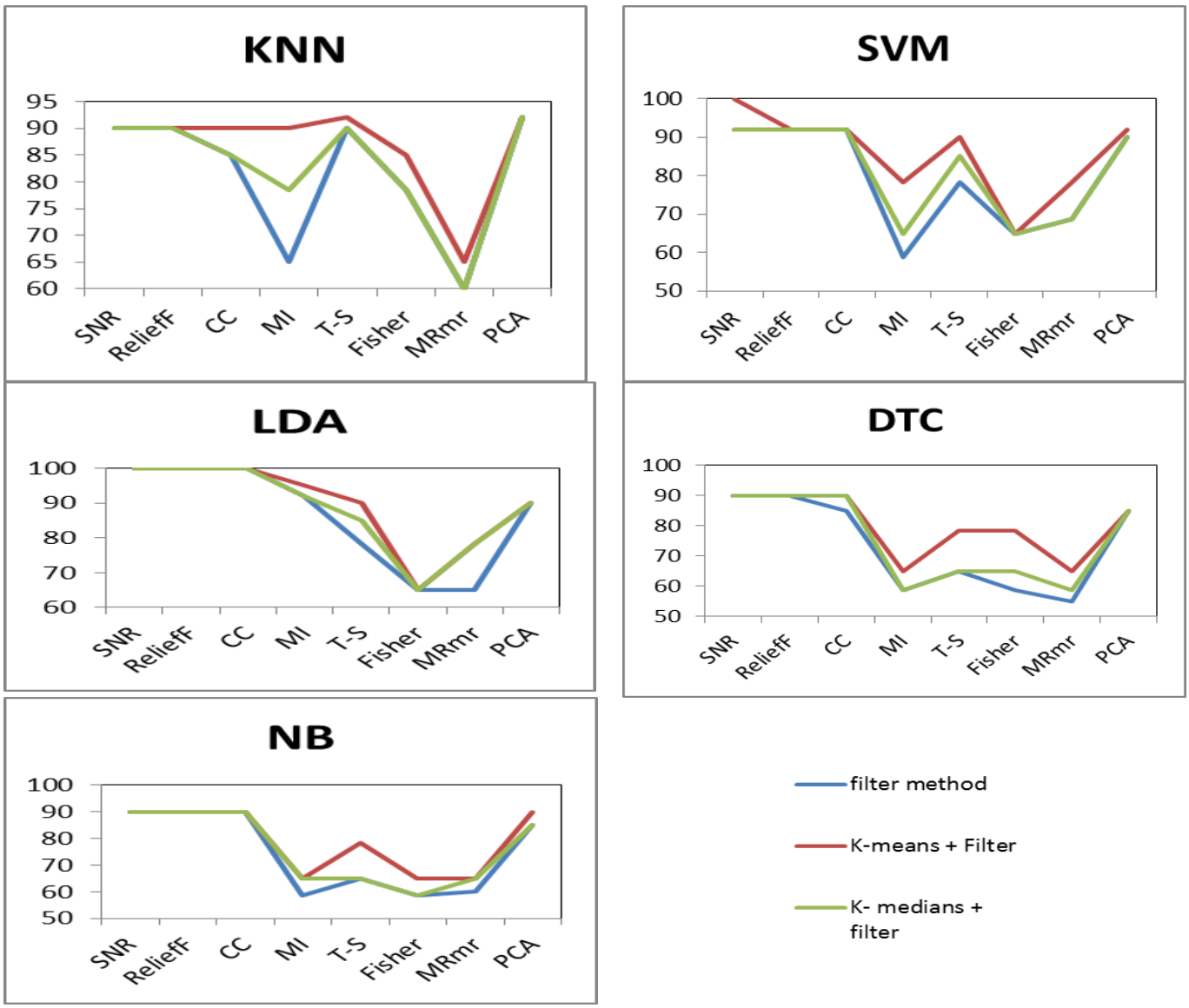

Figure 4. Performance of comparison for proposed classifiers (Prostate cancer)

	KNN		SVM		LDA		DTC		NB	
	Acc (%)	*Nbr genes*	Acc (%)	*Nbr genes*	Acc (%)		Acc (%)	*Nbr genes*	Acc (%)	*Nbr genes*
SNR	100	*6*	100	*33*	100	*64*	97.3	*1*	100	*19*
ReliefF	100	*21*	99.3	*9*	99.3	*80*	97.3	*1*	100	*21*
CC	100	*28*	100	*36*	100	*82*	97.3	*1*	100	*29*
MI	83.2	*10*	88.5	*5*	96.6	*24*	83.2	*7*	96.6	*19*
T-S	99.3	*13*	100	*17*	99.3	*17*	90.6	*1*	96.6	*4*
Fisher	83.2	*82*	88.5	*6*	67.7	*53*	66.4	*3*	84.5	*5*
MRmr	90.6	*62*	88.5	*18*	83.5	*23*	90.6	*5*	90.6	*25*
PCA	99.3	*7*	97.3	*35*	99.3	*64*	99.3	*5*	96.6	*5*
K-means + SNR	**100**	*3*	**100**	*10*	**100**	*14*	99.3	*2*	100	*10*
K-means + ReliefF	**100**	*4*	**100**	*11*	**100**	*28*	99.3	*12*	100	*11*
K-means + CC	**100**	*5*	**100**	*12*	**100**	*19*	99.3	*17*	100	*15*
K-means + MI	**96.6**	*9*	**90.6**	*5*	**99.3**	*20*	90.6	*13*	96.6	*10*
K-means + T-S	99.3	*11*	100	*7*	99.3	*12*	96.6	*5*	99.3	*12*
K-means + Fisher	90.6	*12*	90.6	*15*	88.5	*28*	83.2	*12*	90.6	*18*
K-means + MRmr	90.6	*11*	90.6	*12*	88.5	*13*	96.6	*15*	96.6	*21*
K-means + PCA	99.3	*5*	99.3	*15*	99.3	*6*	99.3	*4*	96.6	*2*
K-medians + SNR	100	*5*	100	*19*	100	*20*	99.3	*9*	100	*15*
K-medians + ReliefF	100	*10*	100	*26*	100	*27*	99.3	*21*	100	*12*
K-medians + CC	100	*13*	100	*23*	100	*23*	99.3	*27*	100	*20*
K-medians + MI	90.6	*17*	90.6	*21*	96.6	*31*	90.6	*17*	96.6	*15*
K-medians + T-S	99.3	*12*	100	*12*	99.3	*15*	96.6	*21*	96.6	*2*
K-medians + Fisher	88.5	*3*	90.6	*19*	88.5	*31*	67.7	*32*	88.5	*16*
K-medians + MRmr	90.6	*34*	90.6	*31*	88.5	*15*	96.6	*21*	90.6	*14*
K-medians + PCA	99.3	*5*	97.3	*12*	99.3	*31*	99.3	*4*	96.6	*3*

Table 5. Performance of comparison for proposed classifiers (lung cancer)

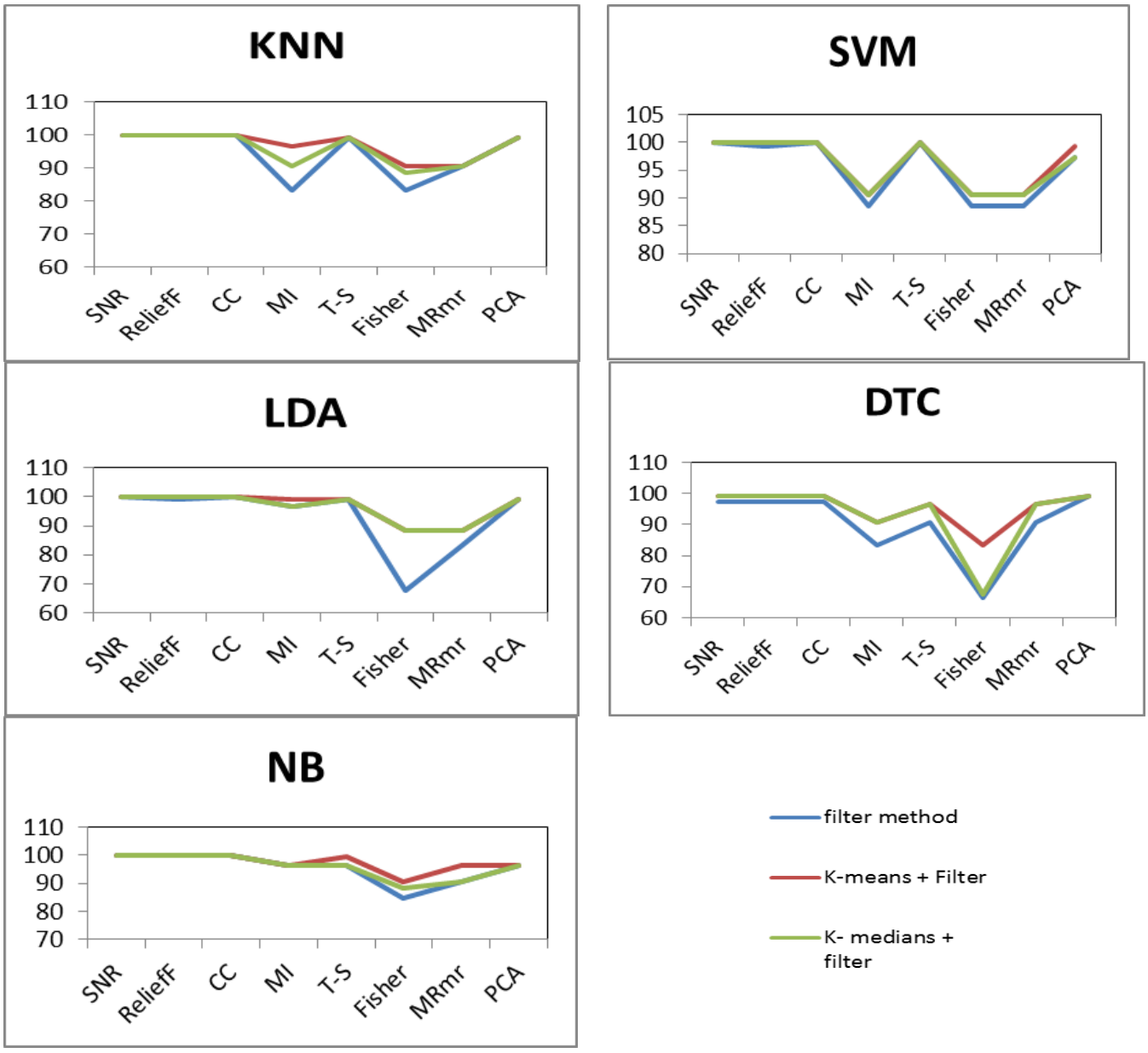

Figure 5. Performance of comparison for proposed classifiers (lung cancer)

	KNN		SVM		LDA		DTC		NB	
	Acc (%)	*Nbr genes*	Acc (%)	*Nbr genes*	Acc (%)		Acc (%)	*Nbr genes*	Acc (%)	*Nbr genes*
SNR	**100**	*4*	**100**	*32*	**100**	*24*	95.6	*3*	95.6	*3*
ReliefF	**100**	*86*	**100**	*20*	**100**	*93*	95.6	*12*	86.9	*2*
CC	**100**	*13*	**100**	*39*	**100**	*97*	95.6	*8*	95.6	*55*
MI	86.9	*10*	86.9	*15*	52.1	*50*	91.3	*13*	91.3	*29*
T-S	91.3	*3*	95.6	*13*	95.6	*4*	95.6	*13*	95.6	*3*
Fisher	86.9	*1*	78.2	*29*	78.2	*1*	82.6	*1*	82.6	*4*
MRmr	86.9	*15*	86.9	*10*	91.3	*5*	91.3	*10*	95.6	*13*
PCA	**100**	*17*	**100**	*43*	**100**	*87*	95.6	*18*	95.6	*25*
K-means + SNR	**100**	*3*	**100**	*10*	**100**	*12*	97	*12*	97	*22*
K-means + ReliefF	**100**	*12*	**100**	*10*	**100**	*17*	95.6	*2*	91.3	*13*
K-means + CC	**100**	*8*	**100**	*4*	**100**	*22*	95.6	*1*	95.6	*12*
K-means + MI	**95.6**	*7*	**97**	*7*	**99.3**	*4*	91.3	*2*	91.3	*3*
K-means + T-S	95.6	*13*	95.6	*3*	97	*2*	95.6	*10*	97	*13*
K-means + Fisher	91.3	*21*	86.9	*32*	86.9	*12*	91.3	*14*	91.3	*15*
K-means + MRmr	91.3	*13*	91.3	*23*	95.6	*16*	91.3	*3*	95.6	*7*
K-means + PCA	100	*8*	100	*7*	100	*38*	97	*12*	97	*15*
K-medians + SNR	100	*3*	100	*28*	100	*19*	97	*21*	95.6	*7*
K-medians + ReliefF	100	*38*	100	*15*	100	*52*	95.6	*7*	91.3	*21*
K-medians + CC	100	*11*	100	*12*	100	*37*	95.6	*5*	95.6	*24*
K-medians + MI	91.3	*3*	91.3	*15*	91.3	*21*	91.3	*5*	91.3	*12*
K-medians + T-S	95.6	*18*	95.6	*10*	97	*23*	95.6	*11*	97	*15*
K-medians + Fisher	91.3	*25*	86.9	*38*	82.6	*14*	91.3	*15*	91.3	*23*
K-medians + MRmr	91.3	*17*	91.3	*35*	95.6	*18*	91.3	*7*	95.6	*11*
K-medians + PCA	100	*12*	100	*31*	100	*52*	95.6	*3*	95.6	*2*

Table 6. Performance of comparison for proposed classifiers (lymphoma cancer)

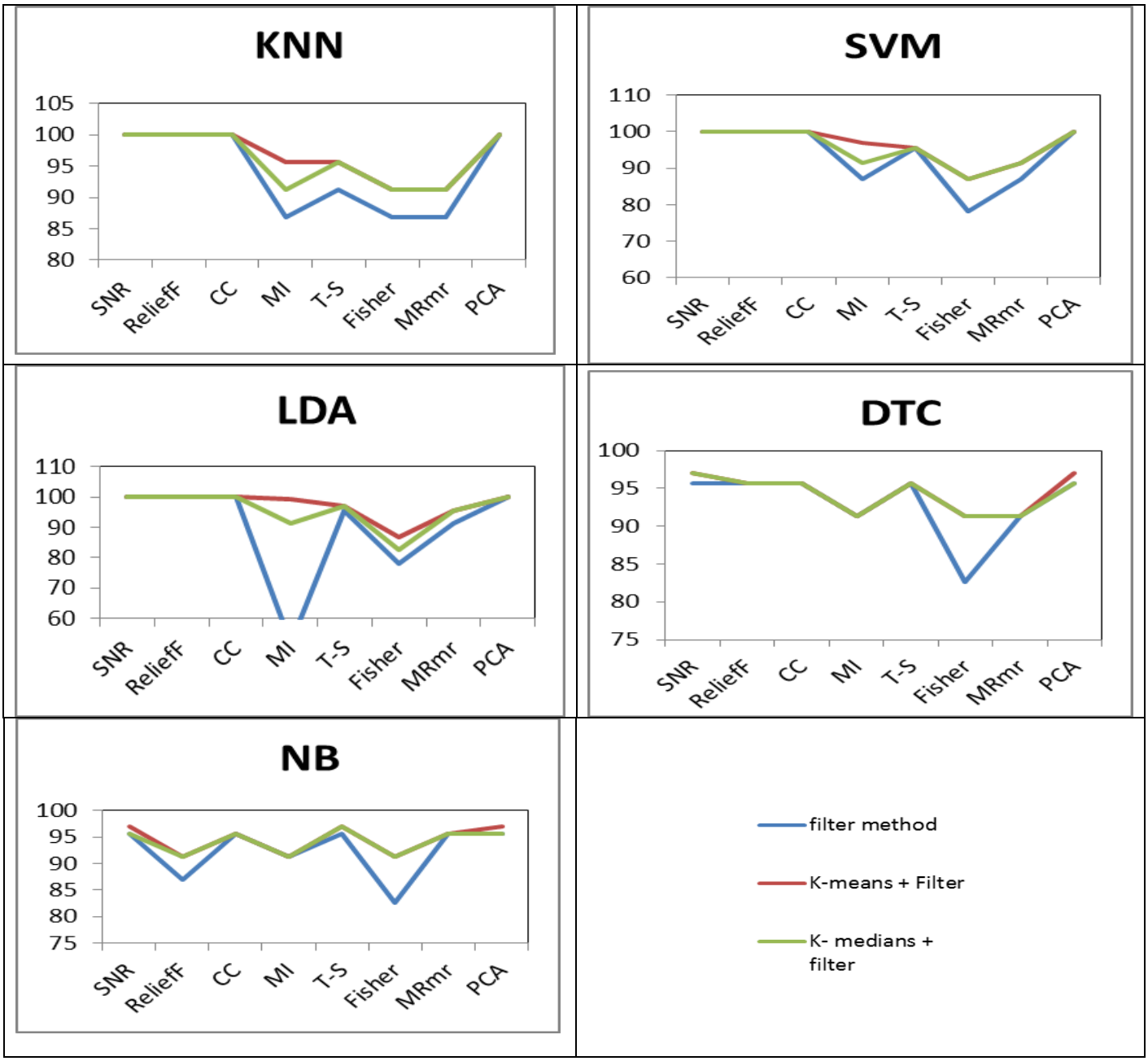

Figure 6. Performance of comparison for proposed classifiers (lymphoma cancer)

	KNN		SVM		LDA		DTC		NB	
	Acc (%)	*Nbr genes*	Acc (%)	*Nbr genes*	Acc (%)		Acc (%)	*Nbr genes*	Acc (%)	*Nbr genes*
SNR	**76.7**	*6*	65.1	*21*	69.7	*28*	58.1	*1*	72	*20*
ReliefF	65.1	*12*	62.7	*48*	62.7	*48*	55.8	*1*	69.7	*13*
CC	72	*13*	65.1	*13*	65.1	*98*	55.8	*1*	72	*12*
MI	65.1	*11*	65.1	*12*	58.1	*23*	55.8	*3*	55.8	*11*
T-S	62.7	*6*	65.1	*20*	62.7	*14*	44.1	*10*	60.4	*2*
Fisher	65.1	*87*	58.1	*67*	69.7	*2*	58.1	*2*	69.7	*31*
MRmr	65.1	*32*	62.1	*13*	62.7	*13*	58.1	*22*	60.4	*13*
PCA	**72**	*13*	**62.7**	*11*	**65.1**	*22*	62.7	*13*	72	*11*
K-means + SNR	**84**	***31***	**72**	*13*	**72**	*18*	69.7	*10*	84	*10*
K-means + ReliefF	**72**	***3***	**72**	*14*	**69.7**	*23*	69.7	*3*	72	*4*
K-means + CC	**84**	***11***	**72**	*12*	**72**	*13*	69.7	*10*	72	*2*
K-means + MI	**69.7**	***3***	**69.7**	*10*	**69.7**	*13*	58.1	*3*	58.1	*6*
K-means + T-S	72	*13*	69.7	*10*	69.7	*3*	58.1	25	62.7	12
K-means + Fisher	72	*19*	69.7	*16*	72	*21*	69.7	*12*	69.7	*3*
K-means + MRmr	69.7	*12*	62.7	*3*	62.7	*2*	62.7	*12*	65.1	*13*
K-means + PCA	84	*35*	72	*12*	69.7	*20*	69.7	*18*	72	*3*
K-medians + SNR	84	*35*	72	*25*	69.7	*3*	69.7	*15*	72	*3*
K-medians + ReliefF	65.1	*3*	62.7	*4*	62.7	*3*	69.7	*10*	69.7	*4*
K-medians + CC	72	*3*	69.7	*10*	65.1	*2*	58.1	*2*	72	*5*
K-medians + MI	69.7	*10*	69.7	*13*	69.7	*21*	58.1	*11*	55.8	*2*
K-medians + T-S	69.7	*3*	65.1	*2*	65.1	*3*	55.8	*13*	62.7	*12*
K-medians + Fisher	69.7	*11*	58.1	*3*	72	*34*	62.7	*3*	69.8	*15*
K-medians + MRmr	65.1	*12*	62.1	*4*	62.7	*11*	62.7	*12*	60.4	*3*
K-medians + PCA	72	*3*	72	*22*	65.1	*3*	62.7	*3*	72	*5*

Table 7**.** Performance of comparison for proposed classifiers (CNS cancer)

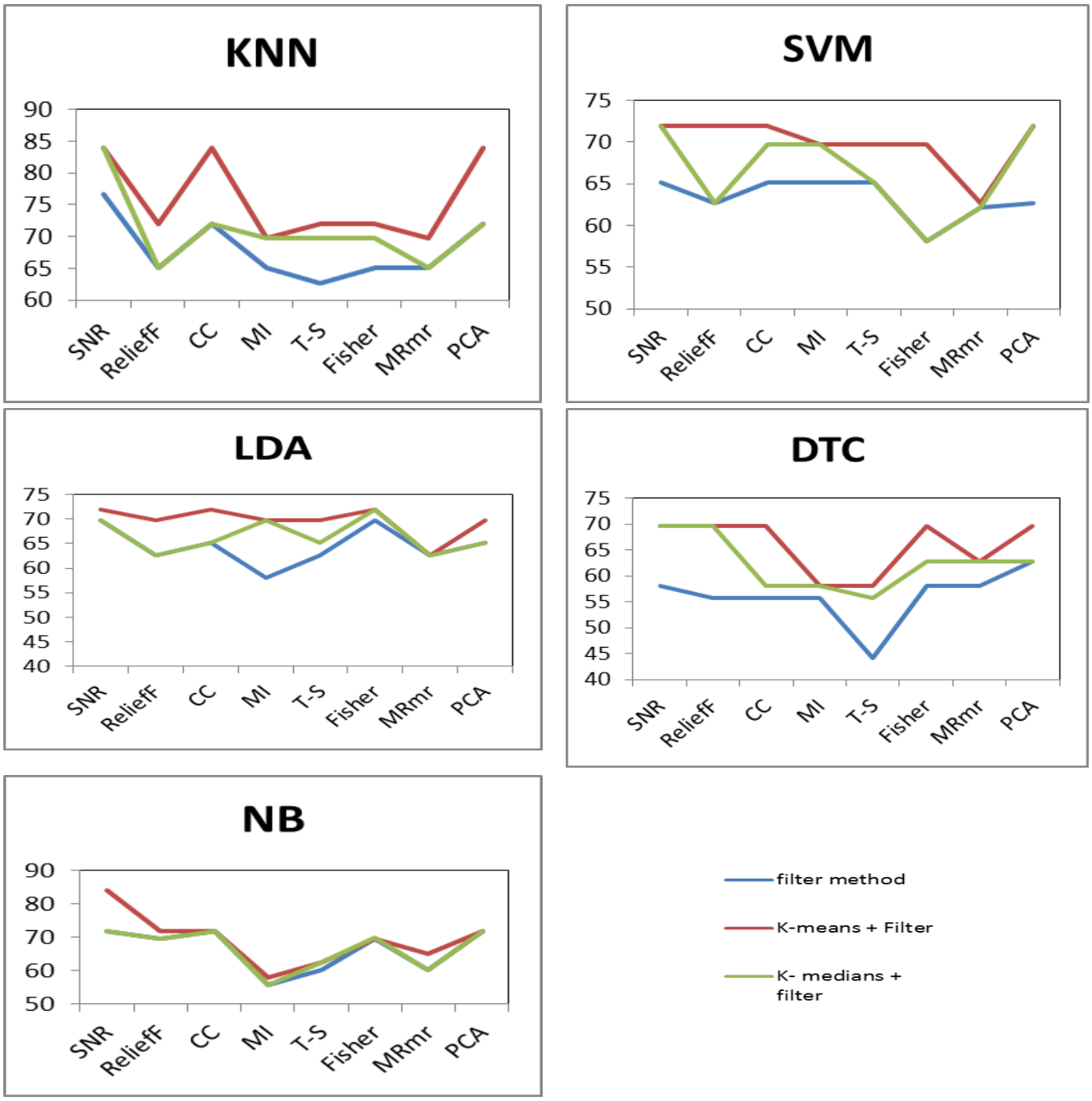

Figure 7. Performance of comparison for proposed classifiers (CNS cancer)

References

[1] Takeshi Saitoh, Toshiki Shibata and Tsubasa Miyazono.Feature Points based Fish Image Recognition. *International Journal of Computer Information Systems and Industrial Management Applications*. Volume 8 (2016) pp. 012–022

[2] Silvia Cateni and Valentina Colla. Improving the stability of wrapper variable selection applied to binary classification. *International Journal of Computer Information Systems and Industrial Management Applications*. Volume 8 (2016) pp. 214–225

[3] Yee Ching Saw, Zeratul Izzah Mohd Yusoh, Azah Kamilah Muda and Ajith Abraham. Ensemble Filter-Embedded Feature Ranking Technique (FEFR) for 3D ATS Drug Molecular Structure. *International Journal of Computer Information Systems and Industrial Management Applications*. Volume 9 (2017) pp. 124-134

[4] Sara Haddou Bouazza, Nezha Hamdi, Abdelouhab Zeroual and Khalid Auhmani. "Gene-expression-based cancer classification through feature selection with KNN and SVM classifiers", 2015 *Intelligent Systems and Computer Vision (ISCV)*, 2015

[5] Ivan Vincent, Ki-Ryong Kwon, Suk-Hwan Lee, Kwang-Seok Moon. Acute lymphoid leukemia classification using two-step neural network classifier. *International Workshop on Frontiers of Computer Vision. IEEE*. MAY 2015.

[6] Logique floue et algorithmes génétiques pour le pré-traitement de données de biopuces et la sélection de gènes, *thèse de doctorat*, edmundobonilla huerta, 2008

[7] Ali El Akadi. Contribution à la sélection des variables pour la classification. *thèse de doctorat*. 2012

[8] Lei Zhang, Yuehui Chen, Ajith Abraham. Hybrid flexible neural tree approach for leukemia cancer classification. *World Congress on Information and Communication Technologies*, 2011

[9] Chanho Park, Sung Bae Cho. Evolutionary ensemble classifier for lymphoma and colon cancer classification. *Conference*: *Evolutionary Computation*, 2003, DOI: 10.1109/CEC.2003.1299385.

[10] Dinesh Singh, Phillip G. Febbo, Kenneth Ross, Donald G. Jackson, Judith Manola, Christine 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*: March 2002, Vol. 1.. Published: 2002.02.28

[11] Gordon GJ, Jensen RV, Hsiao LL, Gullans SR, Blumenstock JE, Ramaswamy S, Richards WG, Sugarbaker DJ, Bueno R: Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. *Cancer Res*. 2002, 62: 4963-4967.

[12] M. A. Shipp, K. N. Ross, P. Tamayo et al., "Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning," *Nature Medicine*, vol. 8, no. 1, pp. 68–74, 2002

[13] Pomeroy, S. L., Tamayo, P. and Gaasenbeek, M. (2002). Prediction of Central Nervous System Embryonal Tumour Outcome Based on Gene Expression. *Nature*, 415, 436–442

[14] Sara Haddou Bouazza, Khalid Auhmani, Abdelouhab Zeroual and Nezha Hamdi. Selecting significant marker genes from microarray data by filter approach for cancer diagnosis. *Procedia Computer Scienc*e 127 (2018) 300–30

[15] Guyon, Isabelle; Elisseeff, André (2003). "An Introduction to Variable and Feature Selection". *JMLR* 3.

[16] MiroslavaCuperlovic-Cuf, Nabil Belacel, Rodney. j. Ouellette, "Determination of Tumour marker genes from gene expression data, Drug Discovery Today", Vol-10, Number 6 pp429-437, 2005

[17] K Kira and L. Rendell. A practical approach to feature selection. *Machine Learning Proceedings*. Page 249-256, 1992.

[18] Robnik-Šikonja, M. & Kononenko. Theoretical and empirical analysis of relieff and rrelieff. Machine Learning, 53(1-2), 23–69.

[19] D. Nguyen and D. Rock. Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics*, 18(1):39–50, 2002.

[20] P. E. H. R. O. Duda and D. G. Stork. Pattern Classification. *Wiley-Interscience Publication*, 2001

[21] Leo Egghe, Lo et Leydesdorff, The relation between Pearson's correlation coefficient r and Salton's cosine measure, *Journal of the American Society for Information Science and Technology*, May, 2009. 10.1002/asi.21009

[22] Ding C, Peng H (2005) Minimum redundancy feature selection from microarray gene expression data. *Journal of Bioinformatics and Computational Biology* 3:185–205

[23] E. Shannon. A mathematical theory of communication. *The bell System Technical Journal*, 27:623–654, 1948.

[24] Akarsh Goyal, Patra Anupam Sourav and Arunkumar Thangavelu. A Comparative Analysis of Simulated Annealing Based Intuitionistic Fuzzy K-Mode Algorithm for Clustering Categorical Data. *International Journal of Computer Information Systems and Industrial Management Applications*. Volume 9 (2017) pp. 232-240

[25] 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*, vol 737. *Springer*, Cham

[26] MacKay, David (2003). "Chapter 20. An Example Inference Task: Clustering". Information Theory, Inference and Learning Algorithms. *Cambridge University Press*. pp. 284–292. ISBN 0-521-64298-1. MR 2012999

[27] Hervé Cardot, Peggy Cénac, Jean-Marie Monnez. A fast and recursive algorithm for clustering large datasets with k-medians. *Computational Statistics and Data Analysis* 56 (2012) 1434–1449

[28] Akram Rajeb, Zied Loukil and Abdelmajid Ben Hamadou. Comparison between two declarative approaches to solve the problem of Pattern Mining in Sequences. *International Journal of Computer Information Systems and Industrial Management Applications.* Volume 8 (2016) pp. 052-056

[29] Alex J. Smola, Bernhard Schölkopf. A tutorial on support vector regression. *Bibliometrics Data Bibliometrics*. August 2004, Volume 14, Issue 3, pp 199-222

[30] Sergey Y. Yurish. Sensors and Biosensors, MEMS Technologies and its Applications. *Advances in Sensors: Reviews*, Vol. 2. Par Sergey Yurish. 2014

[31] Sara haddou bouazza, Khalid auhmani, Abdelouhab 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

[32] Tina R. Patil, Mrs. S. S. Sherekar. Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification. *International Journal Of Computer Science And Applications*. Vol. 6, No.2, Apr 2013

[33] Ayca Çakmak Pehlivanlı. A novel feature selection scheme for high-dimensional data sets: four-Staged Feature Selection. *Journal of Applied Statistics*, 2015