Cancer Classification Using Gene Expression Profiling: Application of the Filter Approach with the Clustering Algorithm

Sara Haddou Bouazza[1](✉), Khalid Auhmani[2], Abdelouhab Zeroual[1], and Nezha Hamdi[1]

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

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

Abstract. In this paper, we investigate the classification accuracy of different cancers based on microarray expression values. For this purpose, we have used hybridization between a filter selection method and a clustering method to select relevant features in each cancer dataset. Our work is carried out in two steps. First, we examine the effect of the filter selection methods on the classification accuracy before clustering. The studied filter selection methods are SNR, ReliefF, Correlation Coefficient and Mutual Information. The K Nearest Neighbor, Support Vector Machine and Linear Discriminant Analyses classifier were used for supervised classification task.

In the second step, the same investigation is carried out, but the feature selection task is preceded by a k-means clustering operation.

Obtained results showed that the best classification accuracies were obtained (for leukemia, colon, prostate, lung and lymphoma cancers datasets) for SNR method. After adding the clustering step to the phase of the feature subset selection, the classification accuracy has been increased for the four selection methods SNR, ReliefF, Correlation Coefficient, and Mutual Information.

Keywords: DNA microarray · Feature selection · Supervised classification
Clustering · Image processing

1 Background

DNA microarrays are characterized by high dimensionality due to the high number of features composed of thousands of genes and a limited number of observations. For this reason, it becomes necessary to reduce the dimensionality of dataset in order to decrease the size of the dataset matrix and also, to make the classification task easier and faster.

One form of the dimensionality reduction is feature subset selection, an imperative step in the field of classification.

So, in order to classify a cancer dataset, we need to select the relevant features that best represent the cancer dataset. To do this, we need to use a filter selection method on

A. Abraham et al. (Eds.): SoCPaR 2017, AISC 737, pp. 90–99, 2018.
https://doi.org/10.1007/978-3-319-76357-6_9

the original cancer dataset, and then use this selected subset of features to classify cancer dataset. The classification accuracy obtained by the classifier represents the performance of the subset selected.

In this paper, we suggest to use the k means clustering not only as a classification algorithm but also as a selection method. We hybridized between a filter selection method (the signal to noise ratio (SNR), ReliefF, Correlation Coefficient (CC), ReliefF and Mutual Information (MI)) and the clustering K-means. To compare these feature selection methods, an evaluation of the dimensionality reduction had been done using four supervised classifiers (k nearest neighbors (KNN), Support Vector Machine (SVM) and Linear Discriminant analysis (LDA)).

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

2 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:

- Fisher, T-statistics, Signal to noise ratio and ReliefF selection methods [1].
- The use of two-step neural network classifier [2].
- The (BW) discriminant score was proposed by [3]. 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) [4].

3 Materials and Methods

We used different feature selection methods and classifiers for cancer classification.

In the first step, we downloaded the dataset of each cancer 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.

3.1 Dataset Description

In this paper, we studied the effect of feature selection methods on three commonly used gene expression datasets: leukemia cancer, Colon cancer and Prostate cancer (Table 1).

- Leukemia is composed of 7129 features and 72 samples. It contains two classes: acute lymphocytic leukemia (ALL) and acute myelogenous leukemia (AML). It can be downloaded from the website[1].

[1] broadinstitute.org/cgi-bin/cancer/publications/pub_paper.cgi?mode = view&paper_id = 43.

Table 1. Datasets and parameters used for experiments

Dataset	No. of features	No. of observation	No. of classes
Leukemia [5]	7129	72	2
Colon [6]	6500	62	2
Prostate [7]	12600	101	2
Lung [8]	12533	181	2
Lymphoma [9]	7070	77	2

- Colon cancer is composed of 6500 features 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 features 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 features 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].

3.2 Feature Subset Selection

Feature selection is the process of selecting a subset of relevant features for model construction (Fig. 1).

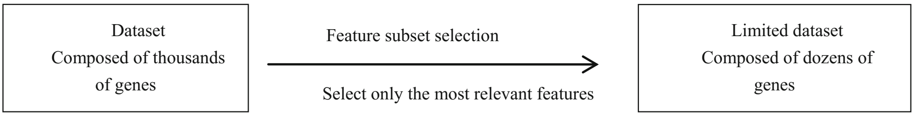

Fig. 1. Feature subset selection

The main idea to apply a feature selection method is that the dataset contains many Features that are either redundant or irrelevant and can consequently be removed without high loss of information.

A feature selection algorithm can be considered as the combination of a search technique for proposing new feature subsets, along with an evaluation measure which scores the different feature subsets. There are three main categories of feature selection algorithms: wrappers, filters and embedded methods [10].

[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.

- 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.

We are interested in this paper with filter methods which are based on the estimated weight (scores) corresponding to each feature (gene) used to order then to select the most relevant descriptors.

The methods used in this work are the Signal to Noise Ratio (SNR), Correlation Coefficient (CC), ReliefF, Mutual Information (MI) and clustering (K-means).

The signal to noise ratio

The signal to noise ratio, called also S/R test, recognizes relevant features by calculating the score S/R of each gene (g) [11].

This score was proposed by [5] and expressed as follows:

$$\mathrm{S/R}_{(\mathrm{g})} = \frac{M_{1g} - \mathrm{M}_{2g}}{\mathrm{S}_{1g} + \mathrm{S}_{2g}} \tag{1}$$

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

ReliefF

This algorithm presented as Relief [12] and then developed and adjusted to the multi-class case by Kononenko as the ReliefF [13].

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:

$$\mathrm{Wd} = \mathrm{wd} - \sum\nolimits_{\mathrm{j}=1}^{\mathrm{K}} \frac{\mathrm{diff}\left(\mathrm{x}_i,\ \mathrm{d},\ \mathrm{hits}_j\right)}{\mathrm{m} * \mathrm{k}} + \sum\nolimits_{\mathrm{c} \neq \mathrm{class}(\mathrm{x}_i)} \frac{\mathrm{p(c)}}{1 - \mathrm{p}(\mathrm{class}(\mathrm{x}_i))} \sum\nolimits_{\mathrm{j}=1}^{\mathrm{k}} \frac{\mathrm{diff}\left(\mathrm{x}_i,\ \mathrm{d},\ \mathrm{misses}_j\right)}{\mathrm{m} * \mathrm{k}} \tag{2}$$

The distance used is defined by:

$$\mathrm{diff}\left(\mathrm{x}_{\mathrm{i}}, \mathrm{d},\ \mathrm{x}_{\mathrm{j}}\right) = \frac{\left|x_{id} - \mathrm{x}_{jd}\right|}{\max(\mathrm{d}) - \min(\mathrm{d})} \tag{3}$$

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. x_{id} is the value of the d_{th} feature of the data x_i.

This method does not eliminate redundancy, but defines a relevant criterion.

Correlation Coefficient

Correlation coefficients measure the strength of association between two features. The Pearson correlation coefficient measures the strength of the linear association between features [14].

Let 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:

$$\rho_{x,y} = \frac{cov(X,\, Y)}{S_x S_y} = \frac{E((X - E(X))(Y - E(Y)))}{S_x S_y} \tag{4}$$

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

Mutual Information

Let us consider a random feature G that can take n values over several measures, we can empirically estimate the probabilities $P(G_1)$, …, $P(G_n)$ for each state G_1, ……, Gn of feature Shannon's entropy [15] of the feature is defined as:

$$\mathrm{H}_{(\mathrm{G})} = -\sum\nolimits_{i=0}^{NG} \mathrm{P}_{(\mathrm{G})} \log\left(\mathrm{P}_{\mathrm{G(i)}}\right) \tag{5}$$

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:

$$\mathrm{MI(G,\, C) = H(G) + H(C) - H(G,\, C)} \tag{6}$$

$$\mathrm{H(G,\, C)} = -\sum\nolimits_{i=0}^{NG} - \sum\nolimits_{j=0}^{NG} \mathrm{P_w(i,\, j)\log(P_w(i,\, j))} \tag{7}$$

Cluster analysis

Cluster analysis or clustering is the task of assembling a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups. In clustering, the k-means algorithm can be used to divide the input data set into k groups or clusters and returns the index of the cluster to which it has assigned each feature.

K-means algorithm is described as follows:

Given an initial set of k means $m_1(1)$,…,$m_k(1)$, the algorithm proceeds by alternating between two steps [16]:

- Assignment step: Assign each feature to the cluster whose mean yields the least within-cluster sum of squares. Since the sum of squares is the squared Euclidean distance, this is intuitively the "nearest" mean

$$S_i^{(t)} = \left\{ x_{p:} \left\| x_p - m_i^{(t)} \right\|^2 \leq \left\| x_p - m_j^{(t)} \right\|^2,\ 1 \leq j \leq k \right\} \tag{8}$$

  Where each x_p is assigned to exactly one $S^{(t)}$, even if it could be assigned to two or more of them.

- Update step: Calculate the new means to be the centroids of the features in the new clusters.

$$m_i^{(t+1)} = \frac{1}{|S_i^t|} \sum\nolimits_{x_j \in S_i^t} x_j \tag{9}$$

3.3 Classification

To compare all feature selection methods, an evaluation of the dimensionality reduction was done using a supervised classification of the three cancers.

Supervised classification is the process of discriminating data, a set of objects or data more widely, so that the objects in the same class are closer to each other than other classes.

To study the performances of the selected features methods, we used the KNN (K nearest neighbors) classifier.

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 $D_E(X, Y)$, where X, Y are samples composed of N features, such that $X = \{X_1, \ldots, X_N\}, Y = \{Y_1, \ldots, Y_N\}$.

$$D_E(X,Y) = \sum\nolimits_{j=1}^{k} \sqrt{\left(X_i^2 - Y_i^2\right)} \tag{10}$$

Support Vector Machines (SVM)

Support vector machines are supervised learning models used for supervised classification [17]. Support Vector Machines are based on two key concepts: the notion of maximum margin and the concept of kernel functions.

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 [18].

To evaluate the performances of the classifiers, we measure the value of the classification accuracy $A_{ccuracy}$ [19]:

$$A_{ccuracy} = 100 * (TP + TN)/(TN + TP + FN + FP) \tag{11}$$

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.

All the algorithms used in this paper have been run using (MATLAB)

4 Results

In this section, we report the results of an experimental study of the effect of the k-means clustering on five commonly used gene expression datasets.

Each dataset is characterized by a group of features, those features are the genes.

After dividing the initial dataset into training data and test data, we applied a subset selection method on training data to select the most relevant features. This subset helps to classify dataset using a classifier (KNN, SVM and LDA). Test data is used to investigate the performances of selection methods and classifiers.

To increase the selection methods performances, we add a clusterisation to the selection step. We divide training data into clusters, and then we select relevant features in each cluster. The obtained subset presents the most relevant features in the dataset.

Tables 2, 3, 4, 5 and 6 compares the classification accuracy obtained (for leukemia, colon, prostate, lung and lymphoma cancers, respectively) before and after adding the k-means 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 four selection methods investigated in this paper.

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

Classifier	KNN				SVM				LDA			
Selection method	Before clustering		After clustering		Before clustering		After clustering		Before clustering		After clustering	
	Max accuracy (%)	Features selected	Max accuracy (%)	Features selected	Max accuracy (%)	Features selected	Max accuracy (%)	Features selected	Max accuracy (%)	Features selected	Max accuracy (%)	Features selected
SNR	100	13	**100**	**5**	97.05	4	**100**	**4**	100	9	**100**	**5**
ReliefF	100	41	**100**	**8**	97.05	2	**100**	**3**	100	69	**100**	**21**
CC	100	50	**100**	**19**	97.05	2	**97.05**	**2**	100	93	**100**	**35**
MI	76.41	56	**91.1**	**18**	84.2	5	**91.1**	**5**	91.1	10	**94.1**	**5**

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

Classifier	KNN				SVM				LDA			
Selection method	Before clustering		After clustering		Before clustering		After clustering		Before clustering		After clustering	
	Max accuracy (%)	Features selected	Max accuracy (%)	Features selected	Max accuracy (%)	Features selected	Max accuracy (%)	Features selected	Max accuracy (%)	Features selected	Max accuracy (%)	Features selected
SNR	92.8	5	**95**	**6**	85.7	29	**100**	**4**	92.8	2	**100**	**8**
ReliefF	85.7	40	**95**	**25**	85.7	11	**92.8**	**7**	78.5	78	**92.8**	**15**
CC	92.8	7	**94.2**	**2**	85.7	2	**95**	**2**	92.8	27	**95**	**14**
MI	85.7	43	**95**	**25**	78.5	5	**91.1**	**5**	71.4	19	**94.1**	**3**

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

Classifier	KNN				SVM				LDA			
Selection method	Before clustering		After clustering		Before clustering		After clustering		Before clustering		After clustering	
	Max accuracy (%)	Features selected	Max accuracy (%)	Features selected	Max accuracy (%)	Features selected	Max accuracy (%)	Features selected	Max accuracy (%)	Features selected	Max accuracy (%)	Features selected
SNR	90	22	**90**	**1**	92	8	**100**	**9**	100	4	**100**	**3**
ReliefF	90	32	**90**	**5**	92	34	**92**	**7**	100	75	**100**	**43**
CC	85	6	**90**	**1**	92	44	**92**	**5**	100	6	**100**	**3**
MI	65	1	**90**	**4**	58.8	56	**78.4**	**10**	92	10	**95**	**8**

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

Classifier	KNN				SVM				LDA			
Selection method	Before clustering		After clustering		Before clustering		After clustering		Before clustering		After clustering	
	Max accuracy (%)	Features selected	Max accuracy (%)	Features selected	Max accuracy (%)	Features selected	Max accuracy (%)	Features selected	Max accuracy (%)	Features selected	Max accuracy (%)	Features selected
SNR	100	6	**100**	**3**	100	33	**100**	**10**	100	64	**100**	**14**
ReliefF	100	21	**100**	**4**	100	17	**100**	**11**	99.3	80	**100**	**28**
CC	100	28	**100**	**5**	100	36	**100**	**12**	100	82	**100**	**19**
MI	83.2	10	**96.6**	**9**	88.5	5	**90.6**	**5**	96.6	24	**99.3**	**20**

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

Classifier	KNN				SVM				LDA			
Selection method	Before clustering		After clustering		Before clustering		After clustering		Before clustering		After clustering	
	Max accuracy (%)	Features selected	Max accuracy (%)	Features selected	Max accuracy (%)	Features selected	Max accuracy (%)	Features selected	Max accuracy (%)	Features selected	Max accuracy (%)	Features selected
SNR	100	4	**100**	**3**	100	32	**100**	**10**	100	24	**100**	**12**
ReliefF	100	86	**100**	**12**	100	2	**100**	**1**	100	93	**100**	**17**
CC	100	13	**100**	**8**	100	39	**100**	**4**	100	97	**100**	**22**
MI	86.9	10	**95.6**	**7**	86.9	15	**97**	**7**	52.1	50	**99.3**	**4**

From the results obtained in Tables 2, 3, 4, 5 and 6 we remark that the k means clustering step obtains a substantial reduction in feature set size maintaining better accuracy compared with results before clustering step, for the chosen Gene datasets of high dimensionality.

5 Conclusion and Discussion

We have shown in this paper that feature selection methods can be applied successfully to a classification situation, using only a limited number of training samples in a high dimensional space of thousands of features.

We performed several studies on leukemia, colon, prostate, lung and lymphoma cancer datasets. The objective was to classify datasets of each cancer into two classes.

The experimental results show that the proposed method has efficient searching strategies and is capable of producing a good classification accuracy with a small and limited number of features simultaneously.

The best result obtained for leukemia cancer is an accuracy of 100% for only 5 genes. For Colon cancer, we obtain 95% for only 6 genes. For prostate cancer, we obtain 90% for 1 gene. For lung cancer, we obtain 100% for 3 genes. For lymphoma cancer, we obtain 100% for 3 genes.

These results encourage adding a clusterisation before the selection step. It increases the classification accuracies and decreases the number of features selected.

References

1. Bouazza, S.H., Hamdi, N., Zeroual, A., Auhmani, K.: Gene-expression-based cancer classification through feature selection with KNN and SVM classifiers. In: 2015 Intelligent Systems and Computer Vision (ISCV) (2015)
2. Vincent, I., Kwon, K.-R., Lee, S.-H., Moon, K.-S.: Acute lymphoid leukemia classification using two-step neural network classifier, May 2015
3. Logique floue et algorithmes génétiques pour le pré-traitement de données de biopuces et la sélection de gènes, thèse de doctorat, edmundobonilla huerta (2008)
4. El Akadi, A.: Contribution to select relevant features in supervised classification: application to the selection of genes for DNA chips and facial characteristics (2012)
5. Zhang, L., Chen, Y., Abraham, A.: Hybrid flexible neural tree approach for leukemia cancer classification. In: World Congress on Information and Communication Technologies (2011)
6. Park, C., Cho, S.B.: Evolutionary ensemble classifier for lymphoma and colon cancer classification. In: Conference: Evolutionary Computation (2003). https://doi.org/10.1109/CEC.2003.1299385
7. Singh, D., Febbo, P.G., Ross, K., Jackson, D.G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A.A., D'Amico, A.V., Richie, J.P., Lander, E.S., Loda, M., Kantoff, P.W., Golub, T.R., Sellers, W.R.: Cancer Cell: March 2002, vol. 1, 28 Feb 2002
8. Gordon, G.J., Jensen, R.V., Hsiao, L.L., Gullans, S.R., Blumenstock, J.E., Ramaswamy, S., Richards, W.G., Sugarbaker, D.J., Bueno, R.: Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. Cancer Res. **62**, 4963–4967 (2002)
9. Shipp, M.A., Ross, K.N., Tamayo, P., et al.: Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nat. Med. **8**(1), 68–74 (2002)
10. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. JMLR **3**, 1157–1182 (2003)
11. Cuperlovic-Cuf, M., Belacel, N., Ouellette, R.J.: Determination of tumour marker genes from gene expression data. DDT **10**(6), 429–437 (2005)

12. Kira, K., Rendell, L.: A practical approach to feature selection, pp. 249–256 (1992)
13. Robnik-Šikonja, M., Kononenko, I.: Theoretical and empirical analysis of relieff and rrelieff. Mach. Learn. **53**(1–2), 23–69 (2003)
14. Egghe, L., Leydesdorff, L.: The relation between Pearson's correlation coefficient r and Salton's cosine measure. J. Am. Soc. Inf. Sci. Technol. **60**, 1027–1036 (2009). https://doi.org/10.1002/asi.21009
15. Shannon, E.: A mathematical theory of communication. Bell Syst. Tech. J. **27**, 623–654 (1948)
16. MacKay, D.: An example inference task: clustering. Information Theory, Inference and Learning Algorithm, pp. 284–292. Cambridge University Press, Cambridge (2003). Chapter 20. ISBN 0-521-64298-1. MR 2012999
17. Smola, A.J., Schölkopf, B.: A tutorial on support vector regression. Stat. Comput. **14**(3), 199–222 (2004)
18. Sergey, Y.: Sensors and biosensors, MEMS technologies and its applications. In: Advances in Sensors: Reviews, vol. 2. Par Sergey Yurish (2014)
19. Pehlivanlı, A.Ç.: A novel feature selection scheme for high-dimensional data sets: four-staged feature selection. J. Appl. Stat. **43**, 1140–1154 (2015)