Prostate Cancer diagnosis based on microarray gene expression profiles

Sara Haddou Bouazza, Abdelouhab Zeroual

(Department of Physics, Faculty of Sciences Semlalia, Cadi Ayyad University, Marrakech, Morocco

Sara.hb.sara@gmail.com, zeroual@uca.ma)

Khalid Auhmani

(Department of Industrial Engineering, National School of Applied Sciences, Cadi Ayyad, Safi, Morocco
k.auhmani@uca.ma)

**Abstract:** With the presence of thousands of genes and only dozens of samples, the dimensionality reduction becomes a necessary task to do. A way of reducing the high number of genes is Feature selection method.

The Feature Selection method permits the analysis of genes expression profiles. It detects and removes the redundant genes, and selects the relevant genes.

In this paper, we present a Prostate cancer Classification based on gene expression profiles. As a first step, we apply a feature selection method on the Prostate cancer dataset, then we select a subset of informative genes for our cancer classification. After that, we train our classifier with the selected subset. And finally, we classify new samples into Tumor samples or Not Tumor samples.

The used Feature Selection methods are Signal to noise ratio, ReliefF, Correlation coefficient, Mutual information, t-Statistics, Fisher, Max-Relevance min redundancy, Principal Component Analysis, Genetic Algorithm, Random Forest, Support Vector Machines and Recursive Feature Elimination.

We classified the Prostate cancer dataset by the use of the K Nearest Neighbor classifier, the Support Vector Machine, the Linear Discriminant Analysis, the Decision Tree for Classification, the Naïve Bayes, and the Artificial Neural Network Classifier.

From the results of Prostate cancer Classification, we deduced that the best Selection and Classification combination for the Prostate cancer is the feature selection method Signal to Noise Ratio with the classifier Linear Discriminant Analysis. We obtain a classification accuracy of 95% for only six genes.

**Keyword:** Prostate cancer, Classification, Feature selection, DNA microarray, dimensionality reduction

1. **Introduction**

Biotechnology is a technology which permits the measurement of the information in human genes, to offer diagnostic tools of all kinds of cancers.

Available datasets are characterized by a high number of genes and a limited number of samples. The obtained dataset needs to be reduced to a limited subset of the relevant genes, and then build an adequate model for cancer classification.

In this paper, we present a comparative study of selection methods for Prostate classification. We will present eleven different selection methods and six classifiers, test on Prostate cancer.

We selected relevant genes by using Signal to noise ratio method (SNR), ReliefF, Correlation coefficient (CC), Mutual information (MI), t-Statistics (t-S), Fisher, Max-Relevance min redundancy (MRmr), Principal Component Analysis (PCA), Genetic Algorithm (GA), Random Forest (RF), a combination of the Support Vector Machines and Recursive Feature Elimination (SVM-RFE). Afterward, we trained the classifier to classify new samples into a Tumor or Not Tumor. For the classification task, we used the K Nearest Neighbor classifier (KNN), Support Vector Machine (SVM), Linear Discriminant Analysis (LDA), Decision Tree for Classification (DTC), Naïve Bayes (NB), and Artificial Neural Network classifier (ANN).

The paper is organized as follows. In Section 2, we propose the definition of the used Feature Selection Methods and classifiers. In section 3, we compare the performances of different feature selection and classification methods for Prostate cancer classification. In section 4 we present results obtained in section 3. Finally, we conclude in Section 5.

1. **Materials and methods**
	1. Feature Selection Methods

With the presence of a limited number of samples and a large number of genes, a learning model tends to overfit, resulting in their performance degenerates [1]. To resolve this problem, dimensionality reduction techniques become an obligatory step which selects a subset of relevant genes in the original dataset without causing any loss of the original information.

Feature Selection methods remove redundant and not informative genes for the studied data set. This step does improve learning performances, and help to build a better model for classification.

There are two known approaches for Feature Selection: Filter and Wrapper methods.

A filter method selects genes independently to the classification algorithm of the dataset, without taking relationships between genes into account. On the opposing, a wrapper method embeds a gene selection method within a classification algorithm.

In this paper, we perform different selection methods. We adopt the filter and wrapper methods to select the relevant genes.

In this section we present a definition of The Signal to noise ratio selection method (SNR), ReliefF, Correlation coefficient (CC), Mutual information (MI), t-Statistics (t-S), Fisher, Max-Relevance min redundancy (MRmr), Principal Component Analysis (PCA), Genetic Algorithm (GA), Random Forest (RF), a combination of the Support Vector Machines and Recursive Feature Elimination (SVM-RFE).

For the classification task, we define the K Nearest Neighbor classifier (KNN), Support Vector Machines (SVMs), Linear Discriminant Analysis (LDA), Decision Tree for Classification (DTC), Naïve Bayes (NB), and the Artificial Neural Network classifier (ANN).

**Fisher**

The Fisher is a statistical test in which the test statistic has an F distribution under the null hypothesis. It gives a score defined as follows [2]:

 F(g) = $\frac{(M1 - M2)² }{ (S1² + S2²) }$ (1)

Where Mk; Sk² denotes the mean and standard deviation of the gene (g) for the class k = 1; 2.

**ReliefF**

The RELIEF is a feature selection algorithm for the binary classification [3], updated in order to improve the reliability of the probability approximation [4].

The weight (W), for each feature (G), is estimated based on the nearest neighbor hit (H) and the nearest neighbor miss (H) of a random instance R.

The process of adjusting weights is done for (m) instances.

The ReliefF algorithm is outlined below:

Set all weights W [G] = 0

**For** i =1 to m **do**

 Randomly select an instance Ri

 Find nearest hot H and nearest miss M

 **For** all G **do**

 W [G] = W[G]-$ \frac{diff(G, R\_{i}, H)}{m}$ + $ \frac{diff(G, R\_{i}, M)}{m}$ (2)

  **end for**

**end for**

The function diff (G, I1, I2) calculates the difference between the values of the feature G for two instances I1 and I2. For nominal features it is defined as diff (G, I1, I2), it is equal to 0 if genotype (G, I1) = genotype (G, I2), otherwise 1.

**T-Statistics (T-S)**

The T Statistics method is a statistic filter selection method which compares two averages and indicates whether or not they are really different from each other in the population from which the groups were sampled. The Score "t" for each gene (g) is used in [5]:

 t(g) = $\frac{X\_{1} -X\_{2}}{\sqrt{ \frac{S\_{1}²}{n\_{1}}+ \frac{S\_{2}²}{n\_{2}}}}$ (3)

Where nk, Xk and Sk² are the size, the average and the variance of classes k = 1, 2.

**Signal to noise ratio (SNR)**

The signal to noise ratio method is a filter selection method proposed by [6], which scores genes and then selects significant genes according to their expression profile [7]. SNR method measures the score as follows:

 S/R (g) = $\frac{M\_{1g} - M\_{2g}}{S\_{1g} + S\_{2g}}$ (4)

Where Mkg, Sk denotes the mean and the standard deviation of the attribute g for samples of class k (class1 and class 2).

**Correlation Coefficient (CC)**

The correlation coefficient is a statistic filter selection method. It calculates the strength and the direction of a linear relationship between two genes [8].

Let $σ\_{x}$ and $σ\_{y}$be the standard deviations of two random genes X and Y, respectively, and cov (X, Y) is the covariance of X and Y. Then the correlation coefficient r between genes is:

$ ρ\_{x,y}$ = $\frac{cov(X,Y)}{σ\_{x}σ\_{y}}$ = $\frac{E(\left(X-E\left(X\right)\right)\left(Y-E\left(Y\right)\right))}{σ\_{x}σ\_{y}}$ (5)

### Mutual Information (MI)

The mutual information is a selection method which measures the dependence between two genes.

For a random gene G(n) that can take n values over several measures, we can empirically estimate the probabilities P(G(n)). Shannon's entropy [9] of the gene is defined as:

G$∑\_{i=0}^{NG}$P (G) log (P G (i)) (6)

The mutual information between one gene G and the class C is measured by the following expression:

 MI (G, C) = H(G) + H(C) - H(G,C) (7)

 H (G, C) = -$∑\_{i=0}^{NG}$-$∑\_{j=0}^{NG}$Pw (i , j) log (Pw (i ,j)) (8)

### Maximum Relevance minimum redundancy (MRmr)

Maximum Relevance minimum redundancy is a feature selection method used to identify characteristics of genes.

Let U ={X1, X2...} denote a set of random genes, C= {c1, c2...} is a distinguished class variable, and S U represent any subset of U.

For this selection method, we should not use genes which are highly correlated among themselves [10]; the redundancy between features should be taken into account.

To measure the redundancy among genes in S we calculate WI(S):

 WI(S) = 1/|S|² ∑ Xi, X j ϵS MI (Xi, X j) (9)

Where (Xi Xj ) is the measure of mutual information between genes Xi and Xj.

Also, the minimum redundancy should be supplemented by the use of a maximum relevance criterion of the genes with respect to the class variable.

To measure the global relevance of the variables in S with respect to C we calculate VI(S):

 VI(S) = 1/|S| ∑ Xi ϵS (C, X j) (8)

To combine redundancy and relevance we use:

 S\* = arg max S⊆U (VI(S) - WI(S)) (9)

**Principal component analysis (PCA)**

The Principal component analysis is a statistical selection method. It reduces the high dimension of a large set of genes to few informative genes. PCA transform a set of correlated genes into a smaller set of uncorrelated genes called principal components [11].

**Genetic Algorithm (GA)**

The Genetic algorithm (GA) solves optimization problems based on a natural selection process which mimics biological evolution.

The GA is an adaptive heuristic search algorithm. At each step, it randomly selects genes from the current population and uses them as parents to produce the children for the next generation. Over successive generations, the population "evolves" toward an optimal solution [12].

**Random Forest**

Random forest is an algorithm for the classification task; it is developed by Leo Breiman [13]. The random forest gene selection method uses both the backward gene elimination (using the Out Of Bag OOB) and the selection based on the importance spectrum.

To construct a classification tree and obtain an OOB error we:

* Make an empty bagger.
* Grow and Train additional trees.
* Add them to the bagger to the ensemble.
* Process inputs for the bagger.
* Prepare the OOB error values.
* Finally, based on the OOB error values, the genes are sorted and ranked.

After fitting all the forests only, the OOB error rates are examined.

**The Support Vector Machine- Recursive Feature Elimination (SVM-RFE)**

The Support Vector Machine- Recursive Feature Elimination selection method is an algorithm composed of a hybrid between the classifier (SVM) and the wrapper selection method (RFE)

The (SVM) is a classification method introduced by [14]. It is based on the notion of maximum margin and the concept of kernel functions.

The selection method (RFE), compute the change in the cost function caused by removing a given gene. This step estimates the effect of removing one feature at a time on the classification accuracy obtained by (SVM).

The (RFE) is an instance of backward feature elimination [15]. Firstly, it trains the classifier (SVM), then computes the ranking criterion for all genes, and finally, removes the gene with the smallest ranking criterion.

* 1. Supervised Classification

Prediction is the main task used in many fields, including machine learning, pattern recognition, signal and image processing etc. [16]

Supervised classification is the task of discriminating data, supervised way so that the objects in the same classes are closer to each other than other groups [17].

We evaluated the Feature Selection outputs by the use of six supervised classifiers: K nearest neighbors (KNN), Support Vector Machine (SVM), Linear Discriminant Analysis (LDA), Decision Tree for Classification (DTC), Naïve Bayes classifier (NB), and Artificial Neural Network classifier (ANN).

**K Nearest Neighbors (KNN)**

K nearest neighbors is an algorithm that stores all available samples and classifies new samples based on a similarity measure.

We can compute the Euclidean distance between two samples using a distance function DE(X, Y), where X, Y are samples composed of N features, such that X = {X1, …, XN }, Y = {Y1, …, YN }.

DE (X, Y) = ∑k j=1 √ (Xi² - Yi²) (10)

**Support Vector Machine (SVM)**

The Support Vector Machine is a supervised classifier. It analyses dataset and recognizes samples [18]. Support Vector Machine algorithm is built on the notion of maximum margin and the kernel functions.

**Linear Discriminant Analysis (LDA)**

Linear Discriminant Analysis is an algorithm used in the machine learning, to search and find a linear combination of genes that characterizes or separates two or more classes of objects [19].

**Decision Tree for Classification (DTC)**

Decision Tree for Classification is a predictive classifier used in data mining and machine learning. The classifier predicts the class of a target sample by learning simple decision rules inferred from the subset of selected genes [20].

**Naïve Bayes (NB)**

The Naive Bayes is a probabilistic supervised classifier used for cancer classification. It uses Bayes theorem and assumes all attributes to be independent given the value of the class variable [21].

**Artificial Neural Network (ANN)**

Artificial Neural Networks (ANN) are parallel and distributed information processing system which consists of a huge number of simple and massively connected processors [22].

KNN, SVM, LDA, DTC, NB, and ANN classifiers performances are evaluated by calculating the classification accuracy (Acc) [23]:

 Acc = 100\* (TP + TN) / (TN + TP+ FN+ FP) (11)

With TP: true positive; FP: false positive; TN: true negative; FN: false negative.

1. **Results**

In this section, we present results obtained by using Matlab to simulate the programs. We used the Prostate cancer dataset to study the performances of the methods used in this paper.

Prostate cancer is a dataset composed of 12600 genes and 102 samples. It contains two classes: Tumor and Not tumor. Data can be found in [24] and can be downloaded from this website: broadinstitute.org/cgi-bin/cancer/publications/pub\_paper.cgi?mode=view&paper\_id=75

We divided the dataset set into training samples and test samples. Training samples are used to select relevant genes and to build models for classification. The test samples are used to test the performances of the selected subset of the relevant genes, and to test the constructed classifiers.

We used eleven filter selection methods: SNR, ReliefF, CC, MI, t-S, Fisher, MRmr, PCA, GA, Random Forest, and SVM-RFE.

We tested the selected genes by six classifiers: KNN, SVM, LDA, DTC, NB, and ANN.

 Table 1 present classification results for Prostate cancer. It shows the number of selected genes by each selection method and the obtained accuracy by each classifier.

Figure 1 presents a comparative curb for eleven selection methods. It permits the evaluation of the strength of each selection method.

Figure 2 compares between six classifiers.

 Table 1: Performance of comparison of proposed classifiers and selection methods for Prostate classification

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | KNN | SVM | LDA | DTC | NB | ANN |
| Acc (%) | Nbrgenes | Acc (%) | Nbrgenes | Acc (%) | Nbrgenes | Acc (%) | Nbrgenes | Acc (%) | Nbrgenes | Acc (%) | Nbrgenes |
| SNR | 91 | 22 | 92 | 8 | **95** | **4** | 92 | 18 | 92 | 22 | 92 | 12 |
| ReliefF | 91 | 32 | 94 | 34 | 94 | 75 | 90 | 42 | 90 | 27 | 92 | 28 |
| CC | 91 | 6 | 94 | 44 | **95** | **6** | 91 | 31 | 91 | 37 | 92 | 18 |
| MI | 65 | 1 | 65 | 56 | 72 | 10 | 58,8 | 12 | 58,8 | 21 | 65 | 16 |
| T-S | 62 | 12 | 78,4 | 31 | 78,4 | 15 | 60 | 31 | 60 | 22 | 65 | 22 |
| Fisher | 60 | 22 | 78,4 | 12 | 78,4 | 22 | 58,8 | 34 | 58,8 | 12 | 58,8 | 13 |
| MRmr | 60 | 7 | 65 | 60 | 65 | 49 | 54,9 | 3 | 58,8 | 23 | 60 | 19 |
| PCA | 90 | 25 | 92 | 18 | 92 | 27 | 85 | 22 | 85 | 15 | 90 | 32 |
| GA | 90 | 12 | 90 | 8 | 90 | 12 | 83 | 13 | 85 | 23 | 90 | 14 |
| Random Forest | 90 | 24 | 93 | 18 | 94 | 6 | 88 | 18 | 90 | 24 | 92 | 14 |
| SVM-RFE | 90 | 16 | 92 | 14 | **95** | **11** | 90 | 22 | 90 | 14 | 90 | 7 |

Figure 1: Feature Selection methods comparison for Prostate cancer classification

Figure 2: Classifiers comparison for Prostate cancer classification

1. **Discussion and conclusion**

The DNA Microarray is a technology which helps to predict cancer classification. It permits the scientific researchers to discover disease mechanisms as well as to explore the genetic reasons for the anomalies occurring in the functioning of the human body.

The selection and classification task are two major problems in the process of Microarray Gene classification. For this reason, it’s mandatory to reduce the dimensionality of the Microarray Gene dataset and to remove noisy and redundant genes. The obtained subset will be used to build a model for classification.

In this paper, we presented a comparative study of eleven selection methods and six classifiers on the Prostate cancer.

As a selection method, we applied the Signal to noise ratio method, ReliefF, Correlation coefficient, Mutual information, t-Statistics, Fisher, Max-Relevance min redundancy, Principal Component Analysis, Genetic Algorithm, Random Forest, and the Support Vector Machines and Recursive Feature Elimination.

For the classification task, we used the K Nearest Neighbor, Support Vector Machine, Linear Discriminant Analysis, and Decision Tree for Classification, Naïve Bayes, and Artificial Neural Network Classifier.

The comparison between the selection methods (Table 1 and Figure 1) shows that SNR, CC, SVM-RFE, ReliefF, Random Forest, PCA, and GA are the strongest selection methods for Prostate cancer. The obtained accuracies are in the range of 95% and 83%.

 The weakest selection methods are MRmr, MI, Fisher, and T-S. The obtained accuracies are in the range of 78.4% and 54.9%.

The strongest classifier between the six classifiers is LDA, then SVM and ANN. The weakest classifiers are NB, DTC, and then KNN (Figure 2).

The adequate combination for Prostate cancer classification is SNR\_LDA. This approach selects the smallest subset of six relevant genes (Table 1). The selected genes are able to train the LDA classifier and offer the highest accuracy which reaches 95%.

**References**

1. Ramón Díaz-Uriarte and Sara Alvarez de Andrés. Gene selection and classification of microarray data using random forest. BMC Bioinformatics. 2006; 7: 3.
2. Miroslava Cuperlovic-Cuf, Nabil Belacel, Rodney. j. Ouellette, “Determination of Tumour marker genes from gene expression data, DDT”, Vol-10, Number 6 pp429-437, 2005
3. K Kira and L. Rendell. A practical approach to feature selection. Page 249-256, 1992.
4. Theoretical and empirical analysis of relieff and rrelieff. Mach. Learn., 53(1-2), 23–69.
5. D. Nguyen and D. Rocke. Tumor classification by partial least squares using microarray gene expression data. Bioinformatics, 18(1):39–50, 2002.
6. Miroslava Cuperlovic-Cuf, Nabil Belacel, Rodney. j. Ouellette, “Determination of Tumour marker genes from gene expression data, DDT”, Vol-10, Number 6 pp429-437, 2005
7. T. Golub, D. Slonim, P. Tamayo, C.Huard,M. Gaasenbeek, J.Mesirov,H. Coller,M. Loh, J.Downing, M. Caligiuri, C. Bloomfield, and E. Lander. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science, 286:531–537, 1999.
8. 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
9. E. Shannon. A mathematical theory of communication. The bell System Technical Journal, 27:623–654, 1948.
10. Ding C, Peng H (2005) Minimum redundancy feature selection from microarray gene expression data. Journal of Bioinformatics and Computational Biology 3:185–205
11. Raychaudhuri,S., Stuart,J.M. and Altman,R.B. (2000) Principal components analysis to summarize microarray experiments: application to sporulation time series. Pac. Symp. Biocomput., 455–466.
12. Yacci, Paul, "Feature selection of microarray data using genetic algorithms and artificial neural networks" (2009). Thesis. Rochester Institute of Technology
13. L. Breiman, 2001, “Random forests”, Machine Learning, 45:5-32.
14. B. E. Boser, I. M. Guyon, and V. N. Vapnik, 1992, “A training algorithm for optimal margin classifiers”, In D. Haussler, editor, 5th Annual ACM Workshop on COLT, pages 144–152, Pittsburgh, PA.
15. R. Kohavi and G. John, 1997, “Wrappers for feature subset selection”, Artificial Intelligence, 97:12, 273–324.
16. Franck Rapaport, Andrei Zinovyev, Marie Dutreix, Emmanuel Barillot, and Jean-Philippe Vert. Classification of microarray data using gene networks. BMC Bioinformatics. 2007; 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. [Alex J. Smola](http://link.springer.com/search?facet-author=%22Alex+J.+Smola%22),  [Bernhard Schölkopf](http://link.springer.com/search?facet-author=%22Bernhard+Sch%C3%B6lkopf%22). A tutorial on support vector regression. August 2004, Volume 14, [Issue 3](http://link.springer.com/journal/11222/14/3/page/1), pp 199-222
19. Sergey Y. Yurish. Sensors and Biosensors, MEMS Technologies and its Applications. Advances in Sensors: Reviews, Vol. 2. Par Sergey Yurish. 2014
20. M. B. A. Snousy, H. M. El-Deeb, K. Badran, and I. A. A. Khlil, “Suite of decision tree-based classification algorithms on cancer gene expression data,” Egyptian Informatics Journal, vol. 12, no. 2, pp. 73–82, 2011.
21. 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
22. [14] P.Ganesh Kumar and D. Devaraj, 2010, “Intrusion Detection using Artificial Neural Network with Reduced Input Features”, ICTACT Journal on Soft Computing, Issue:01, pp: 30-36.
23. Tom Fawcett. ROC Graphs: Notes and Practical Considerations for Researchers. March 16, 2004
24. 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: March 2002, Vol. 1.. Published: 2002.02.28