information

MDPI

Article

An Experimental Comparison of Feature-Selection and Classification Methods for Microarray Datasets

Nicole Dalia Cilia, Claudio De Stefano *, Francesco Fontanella, Stefano Raimondo and Alessandra Scotto di Freca

Department of Electrical and Information Engineering "Maurizio Scarano", University of Cassino and Southern Lazio, 03043 Cassino (FR), Italy; nicoledalia.cilia@unicas.it (N.D.C.); fontanella@unicas.it (F.F.); stefano.raimondo@unicas.it (S.R.); a.scotto@unicas.it (A.S.d.F.)

* Correspondence: destefano@unicas.it

Received: 29 January 2019; Accepted: 5 March 2019; Published: 10 March 2019

Abstract: In the last decade, there has been a growing scientific interest in the analysis of DNA microarray datasets, which have been widely used in basic and translational cancer research. The application fields include both the identification of oncological subjects, separating them from the healthy ones, and the classification of different types of cancer. Since DNA microarray experiments typically generate a very large number of features for a limited number of patients, the classification task is very complex and typically requires the application of a feature-selection process to reduce the complexity of the feature space and to identify a subset of distinctive features. In this framework, there are no standard state-of-the-art results generally accepted by the scientific community and, therefore, it is difficult to decide which approach to use for obtaining satisfactory results in the general case. Based on these considerations, the aim of the present work is to provide a large experimental comparison for evaluating the effect of the feature-selection process applied to different classification schemes. For comparison purposes, we considered both ranking-based feature-selection techniques and state-of-the-art feature-selection methods. The experiments provide a broad overview of the results obtainable on standard microarray datasets with different characteristics in terms of both the number of features and the number of patients.

Keywords: feature-selection; feature ranking; classification methods; DNA microarrays

1. Introduction

In the last decade, there has been a growing scientific interest in the analysis of DNA microarray datasets, which involved different research fields, such as biology, bioinformatics, and machine learning. DNA microarrays contain measures relative to tests for quantifying the types and the amounts of Messenger Ribonucleic Acid (mRNA) transcripts present in a collection of cells, thus allowing researchers to characterize tissue and cell samples regarding gene expression differences. These data have been widely used in basic and translational cancer research. In this context, DNA microarrays were used either to separate healthy patients from oncological ones, based on their "gene expression" profile (two class problems), or to distinguish between different types of cancer (multiclass problems).

DNA microarray experiments typically generate a very large number of features: for each patient, in fact, each feature represents a gene expression coefficient corresponding to the abundance of Messenger Ribonucleic Acid (mRNA) in a concrete sample [1]. The effect is that the number of features in the raw data ranges from 6000 to 60,000, while the available datasets usually refer to less than 100 patients [2]. This implies that the classification task is very challenging, due to both the high number of gene expression and the small number of samples [3], and require the implementation of an accurate

feature-selection process to reduce the complexity of the feature space and to identify a subset of highly distinctive genes [4,5]. It has been demonstrated, in fact, that most of the genes measured in DNA microarray experiments are not relevant to improve classification accuracy for many of the classes to be recognized [6]. The application of a feature-selection process, however, is not only due to the need for the elimination of redundant or non-distinctive features, but also to more accurately correlate gene expression to diseases. This last aspect is very important for biologists and explains the scientific interest for the development of new features selection algorithms, as evidenced by the large number of works published in the literature in recent years. Nevertheless, it has been highlighted that in this framework there are no standard state-of-the-art results generally accepted by the scientific community: therefore, it is difficult to compare the effectiveness of new algorithms and decide which approach can provide better results in the general case. The problem is even more complex when considering the large number of microarray data sets available whose properties, in terms of both the number of features and the number of patients, can vary significantly [2].

Based on these considerations, the aim of the present work is to provide a broad experimental comparison on the feature-selection and classification techniques applied to different DNA microarray datasets. Unlike other works that have analyzed these aspects separately [2,7], our aim is to study the combined effects of the feature-selection process on the classification results, and to evaluate the sensitivity of the considered classification methods with respect to the size of the feature space. Moreover, taking into account the complexity of the feature-selection problem in case of thousands of features, we focused our analysis on standard feature ranking techniques and we evaluated the classification methods considering in the experiments feature sets of increasing size, obtained by selecting the features according to the order in which they appear in the ranking [8]. For the sake of completeness, we have also compared the obtained results with those relative to the use of three state-of-the-art feature-selection methods.

In the experiments, we considered six DNA microarray datasets with different properties in terms of both the number of features and the number of patients. As for the classification process, we chose four classification methods, namely decision trees, random forest, nearest neighbor and multilayer perceptron, thus providing the interested reader with a broad overview of the results obtainable with the most efficient and widely used classification schemes.

The remainder of the paper is organized as follows: Section 2 briefly illustrate the characteristics of the considered DNA microarray datasets, Section 3 discusses the feature-selection methods, while Section 4 presents the experimental results. Some conclusions are eventually left to Section 5.

2. The Considered DNA Microarray Datasets

We tested the approaches considered for the experimental comparison on the following, publicly available, microarray datasets: *Breast*, *Colon*, *Leukemia*, *Lymphoma*, *Lung* and *Ovarian*. The characteristics of the datasets are summarized in Table 1. They differ in terms of total number of available attributes, number of classes, and number of samples.

Let us denote with N_S the number of samples and with N_G the number of genes. As anticipated in the Introduction, the samples in each dataset represent expression levels of genes in cancer or normal (healthy) tissues. It is worth remarking that for all the datasets $N_G >> N_S$ (N_G is of the order of thousands, while N_S of the hundreds). These datasets are described in the following.

- **Breast** cancer typically develops in either the lobules or the ducts of the breast, but it can also occur in the fatty tissue or the fibrous connective tissue within the breast. The samples of the Breast dataset were extracted from frozen tumor tissue samples from patients with lymph-node-negative breast cancer, but who did not receive systemic neo-adjuvant or adjuvant therapy [9]. The dataset contains 286 samples, labelled according to the estrogen receptors (ER), with 219 of them below the cut-off value and the remaining ones above this value. As concerns the number of genes, each sample is represented by 17,816 genes.

- **Colon** tumor is a disease in which cancerous growths (tumors) are found in the tissues of the colon. This dataset contains 62 samples. Among them, 40 tumor biopsies are from colon adenocarcinoma specimens (snap-frozen in liquid nitrogen within 20 min of removal), whereas 22 normal biopsies are from healthy parts of the colons of the same patients. The total number of genes to be tested is 2000. The samples were analyzed with an Affymetrix oligonucleotide array [10].
- **Leukemia** is a primary disorder of bone marrow. They are malignant neoplasms of hematopoietic stem cells. The Leukemia dataset contains 72 samples, each represented by 7129 genes. All samples represent acute leukemia patients, either acute lymphoblastic leukemia (ALL) or acute myelogenous leukemia (AML) [6].
- The term **lymphoma** encompasses a broad variety of cancers of the lymphatic system, and the diffuse large B-cell lymphoma (DLBCL) is the most common subtype of non-Hodgkin's lymphoma. The Lymphoma dataset contains samples representing DNA microarrays of gene expression in B-cell malignancies [11]. The total number of available genes is 4026 and the number of samples is 62. As concerns the number classes, the samples belong to nine different classes, each representing a different stage of the disease. The DNA microarray analysis of gene expression was done as described in [12].
- **Lung** cancer is one of the most common and harmful cancer and it is characterized by uncontrolled cell growth in the lung tissues. Using established methods, distinguishing between malignant pleural mesothelioma (MPM) and adenocarcinoma (ADCA) of the lung can be cumbersome. However, techniques based on the expression levels of a small number of genes, can be useful in the early and accurate diagnosis of MPM and lung cancer. The lung dataset contains 181 tissue samples (31 MPM and 150 ADCA), each described by 12,533 genes [13].
- **Ovarian** cancer is caused the by the uncontrollable growth of the cells in the ovaries. This growth produces a lump of tissue. It is one of the most common types of cancer in women and mainly affects those over the age of 50. Moreover, ovarian cancer symptoms are similar to those of some more common conditions, and this make it not always easy to diagnose. The Ovarian dataset contains 216 samples (100 patients and 116 controls). In this dataset, all major epithelial subtypes of ovarian cancer are represented [14].

Table 1. The datasets considered for the experiments.

Datasets	Attributes	Samples	Classes
Breast	17,816	286	2
Colon	2000	62	2
Leukemia	7129	72	2
Lymphoma	4026	96	9
Lung	12,533	181	2
Ovarian	2190	216	2

3. Feature Selection

Feature selection is the process of reducing the dimensionality of the available data, with the aim of improving the recognition results. This process typically consists of three steps: a search procedure for searching the solution space made of all the possible solutions, i.e., feature subsets, an evaluation function, and a stopping criterion.

The main difficulty of feature selection is due to the dimension of the search space, which grows exponentially with the number of available features: for data represented by N features, there are 2^N feasible solutions. This makes the exhaustive search strategy computationally intractable [15]. This is typically the case of microarray datasets in which, as previously evidenced, N is of the order of thousands. To overcome this problem, heuristic algorithms are typically used for finding near-optimal solutions.

As for the evaluation functions, they are generally subdivided into three wide categories, namely filter, wrapper and embedded methods [16,17]. Filter methods measure statistical or geometrical properties of the subset to be evaluated, whereas wrapper functions adopt as evaluation measure the accuracy achieved by a given, previously chosen, classifier. Finally, embedded approaches include feature selection in the training process, thus reducing the computational costs due to the classification process needed for each subset.

As just mentioned, wrapper methods are computationally costly because the evaluation of each subset requires the training of the adopted classifier. For this reason, they are typically used with near-optimal search strategies, which can achieve acceptable results, but limiting the computational costs. As for the filter methods, they need non-iterative computations on the dataset which are, in most of the cases, significantly faster than classifier training sessions. Moreover, filter approaches, estimating intrinsic properties of the data, provide more general solutions, typically performing well on a wide family of classifiers.

A particular category of filter methods is that of the ranking ones, which evaluate each feature singularly. Once all features have been evaluated, they are ranked according to their merit. Then the subset search step is straightforward: the best M features are selected, with M set by the user. Unfortunately, even if this approach is very fast and allow dealing with thousands of features, there is no one general criterion for choosing the dimension of the feature space, then it is difficult to select the number M of features to be selected. Moreover, most importantly, relevant features that are highly informative when combined with other ones could be discarded because they are weakly correlated with the target class.

In this study, because of the huge dimensionality of the datasets used for the experiments, we have considered only standard ranking algorithms to build up an experimental protocol for selecting the feature subsets allowing us to achieve the best classification results. The adopted protocol did not use any search strategy, but adds the features progressively, according to their position in the ranking. As for the ranking, we have considered five standard univariate measures, namely Chi-square, Relief, Gain Ratio, Information Gain, and Symmetrical Uncertainty. These measures, as well as the state-of-the-art feature-selection approaches considered for the comparison, are detailed in the following subsections.

3.1. Chi-Square

The Chi-Square (CS) approach implements a discretization algorithm based on the CS statistic [18]. For each feature, the values present in available data are first sorted, and each value represent an interval. Then the Chi-square statistic is used to assess the class relative frequencies of adjacent intervals. For each couple of adjacent intervals, if their frequency similarities are above a given threshold, they are merged. Class similarities are computed according to the following formula [18]:

$$CS = \sum_{i=1}^{2} \sum_{j=1}^{C} \frac{\left(A_{ij} - E_{ij}\right)^2}{E_{ij}} \tag{1}$$

where C is the number of classes, A_{ij} is the number of instances of the j-th class in the i-th interval and E_{ij} is the expected frequency of A_{ij} given by the formula:

$$E_{ij} = R_i C_i / NT \tag{2}$$

where R_i is the cardinality of the i-th interval and C_j and NT represent the number of samples of the j-th class and the total number of samples, respectively, in the two adjacent intervals. The stopping criterion of the merging process is based on the choice of a threshold, which represents the maximum value of the frequency differences in adjacent intervals. In the experiments detailed in Section 4, this value was set according to the results of some preliminary experiments.

3.2. Relief

The Relief (*Re*) measure uses an instance-based learning approach to assign a weight to each feature, representing its relevance with respect to the target concept [19]. This weight is computed by finding for each sample the nearest neighbor of the same class (nearest hit) and the nearest neighbor of a different class (nearest miss). A given feature receives a high weight if it takes different values for instances from different classes and similar values for instances belonging to the same class. Given a feature X, whose values belong to the set of discrete values $\{x_1, x_2, \ldots, x_n\}$, its weight is computed according to the following formula [19]:

$$Re(X) = \frac{I_G(X) \sum_{x_i \in X} p(x_i)^2}{(1 - \sum_{c_k \in S_C} p(c_k)^2) \sum_{c_k \in S_C} p(c_k)^2} \tag{3}$$

where $p(x_i)$ is the probability that the feature X assumes the value x_i, S_C is the set of classes to be discriminated, $p(c_k)$ is the a-priori probability of the k-th class and I_G is a modified version of the Gini index [19], given by the following formula:

$$I_G(X) = \sum_{x_i \in X} \left(\frac{p(x_i)^2}{\sum_{x_j \in X} p(x_j)^2} \sum_{c_k \in S_C} p(c_k x_i)^2 \right) - \sum_{c_k \in S_C} p(c_k)^2 \tag{4}$$

3.3. Information Gain, Gain Ratio Gain and Symmetrical Uncertainty

The measures detailed in this subsection are based on the well-known information-theory concept of entropy. Given a discrete variable X, which can take the values $\{x_1, x_2, \ldots, x_n\}$, its entropy $H(X)$ is defined as:

$$H(X) = -\sum_{i=1}^{n} p(x_i) log_2 p(x_i) \tag{5}$$

where $p(x_i)$ is the probability mass function of the value x_i. In practice, $H(X)$ estimates the uncertainty of the random variable X.

The conditional entropy of two discrete random variables X and Y, taking the values $\{x_1, x_2, \ldots, x_n\}$ and $\{y_1, y_2, \ldots, y_n\}$ respectively, is defined as:

$$H(X|Y) = -\sum_{i=1}^{n} \sum_{j=1}^{n} p(x_i, y_j) log_2 \frac{p(x_j)}{p(x_i, y_j)} \tag{6}$$

where $p(x_i, y_j)$ is the joint probability that $X = x_i$ and $Y = y_j$. In this case, $H(X|Y)$ represents the amount of randomness in the random variable X when the value of Y is known.

The concepts of entropy and conditional entropy just defined can be used to evaluate the effectiveness of a given feature in predicting the class of unknown samples. The first of the three information-theory-based measures considered in this study is denoted as *Information Gain* (*IG*) [20]. For a given a feature X, $IG(X)$ is computed in terms of both the class entropy $H(C)$ and the conditional entropy $H(C|X)$. More specifically, such quantities are used to compute $IG(X)$ as follows:

$$IG(X) = H(C) - H(C|X) \tag{7}$$

It is worth noting that $IG(X)$ measures the reduction of the randomness of the variable C when X is known. In practice, $IG(X)$ measures the quantity of information about C provided by the feature X.

Finally, the Gain Ratio and Symmetrical Uncertainty measures are based on the just mentioned IG measure. More specifically, Gain Ratio (GR) is computed according to the following formula [21]:

$$GR(X) = \frac{IG(X)}{H(X)} \tag{8}$$

As for the Symmetrical Uncertainty (SU), instead, it has been defined in such a way to compensate the bias toward the attributes taking more values. This is achieved by normalizing its value to the range $[0, 1]$ [20]:

$$SU(X) = 2.0 \frac{IG(X)}{H(C) + H(X)} \tag{9}$$

3.4. Heuristic Search Algorithms

For the sake of comparison, we have also considered three state-of-the-art heuristic search algorithms, namely the *Sequential Forward Floating Search*, the *Fast Correlation-Based Filter* and the *Minimum Redundancy Maximum Relevance*, whose properties are briefly summarized in the following.

- *Sequential Forward Floating Search*. This feature-selection algorithm is based on a greedy hill-climbing iterative strategy. Given a feature subset evaluation function $f(S)$, it starts with the empty set S_0 and, at the i-th step, selects among the available features (i.e., those not selected yet) the feature X_b that produces the highest increment of $f(S_{i-1} \cup X_b)$. Once X_b has been selected, the algorithm also checks if $f(S_i)$ can be increased excluding from S_i one of the features previously selected.
  It is worth noting that we used an improved version of this algorithm, presented in [22]. It will be denoted as *SFFS* in the following.
- *Fast Correlation-Based Filter*. This algorithm searches for subsets containing the features that are most relevant to the target concept and least redundant each other. This search is driven by an evaluation function that is based on the concept of mutual information I. Given two random variables X and Y, the mutual information $I(X, Y)$ quantifies the amount of information obtained about one random variable through observing the other random variable. The mutual information is first used to sort the available features, in decreasing order, according to mutual information with the class. Once the features have been ranked, the final subset is built by iteratively selecting the top-ranked features, but skipping the redundant ones, i.e., those that have a high mutual information level with one of the previously selected features [23]. It will be denoted as *FCBF* in the following.
- *Minimum Redundancy Maximum Relevance*. This approach searches for subsets that maximize the relevance with the target class and minimize the redundancy among the selected features. Given a feature subset S, its relevance is computed by averaging the mutual information $I(x_i, C)$, with $x_i \in S$ and C representing a class. As for the redundancy of S, it is computed by averaging $I(x_i, x_j)$, with $x_i, x_j \in S$. It is worth noting that the mutual information can be computed for both discrete and continuous features. In case of continuous features, this quantity is obtained by using the Parzen windows method to estimate feature probability densities. Further details of the algorithm can be found in [24]. It will be denoted as *MRMR* in the following.

4. Experimental Results

As anticipated in the Introduction, the effectiveness of the feature subsets obtained by applying the selection techniques described in Section 3, has been evaluated by using four different classification schemes, namely decision trees, random forest, nearest neighbor, and multilayer perceptron. For each of them, we performed a set of experiments using the microarray datasets described in Section 2, following the same experimental protocol. The adopted classification schemes, the experimental protocol, and the sets of experiments are described in the following subsections.

4.1. The Classification Schemes

As for the classifiers considered in this study, we used the implementation provided by the WEKA tool [25] choosing a 10-fold cross validation strategy. Furthermore, to manage the randomness of

classification algorithms, we performed 20 runs for each experiment. All the results reported below were obtained by averaging the values over the 20 runs.

A *decision tree* (DT) is a classification method that uses a knowledge representation based on a tree structure. In such structure, the internal nodes represent specific tests on the features used to represent patterns, while the branches coming out of a node represent the possible results of these tests. Finally, the leaf nodes represent the class assigned to an input sample based on the results of all the tests performed along a path from the root node to the leaf node. In this study, we considered C4.5 learning algorithm, which builds the decision tree with a top-down approach, using the concept of information entropy. In practice, given a training set TS, it breaks down TS into smaller subsets by gradually increasing the depth of the tree, until it reaches a leaf node in which the process terminates. In each node of the tree, C4.5 chooses the feature that most effectively allows the splitting of the corresponding sample subsets, using the normalized entropy gain as a splitting criterion: this criterion measures how much the obtained subsets are homogeneous, in terms of class labels.

The term *random forest* (RF) does not refer to a single algorithm, but rather to a family of methods for creating an ensemble of tree-based classifiers. The original algorithm proposed by Breiman in [26] is usually denoted *Forest-RI* in the literature and it is considered a reference method in most of the papers dealing with RF. Given a training set that contains N_f feature vectors, each consisting of N features, the forest-RI algorithm is composed of the following steps applied to each tree:

1. Randomly extract N_f samples with replacement from the dataset. The obtained set of samples will be used as training set for the starting node of the tree.
2. Set a number $K \ll N$.
3. At each node, randomly select K features from the whole set of available ones.
4. For each of the selected features, consider its values in the training set and choose the best binary split value according to the Gini index [27]. Then, choose the feature with the best index value and generate two new nodes by splitting the samples associated with the original node according to such a value.
5. Increase the depth of the tree to its maximum size, according to the defined stopping criterion. Note that node splitting usually is stopped when one of the following conditions occur:
   (i) The number of samples in the node to be split is below a given threshold;
   (ii) all the samples in the node belong to the same class.
6. Leave the tree not pruned.

Once the forest has been built, an unknown sample is labeled according to the Majority Vote rule: i.e., it is labeled with the most popular class among those provided by the ensemble trees.

The *K-Nearest Neighbor* algorithm (K-NN) is a well-known non-parametric classification technique, according to which an unknown sample is assigned the most frequent class among those of its k-nearest samples of the training set. The rationale behind this technique is that given an unknown sample $\mathbf{x}$ to be assigned to one of the c_i classes of the considered application field, the a-posteriori probabilities $p(c_i\mathbf{x})$ in the neighborhood of $\mathbf{x}$ may be estimated by taking into account the class labels of the k-nearest neighbors of $\mathbf{x}$. Although its simplicity, K-NN proved to be able to provide high performance. The results shown below were obtained by using the Mahalanobis distance, which demonstrated to be more effective than the Euclidean distance in a set of preliminary experiments. As regards the value of the parameter k, we found that the value $k = 1$ significantly outperformed the higher values. This result suggests that for these kinds of data, higher value of k force the algorithm to consider far neighbors belonging to classes different from the actual class of the sample at hand. The effect is that of assigning several samples to a wrong class, reducing the overall recognition rate. This behavior could also be a consequence of the limited number of specimens available in the training set, as well as a specificity of the considered DNA microarray datasets.

An *artificial Neural Network* (NN in the following) is a well-known information processing paradigm inspired by the way biological nervous systems process information. These networks typically consist of many highly interconnected processing elements (neurons), often equipped with a local memory, that can process information locally, providing as output a single unidirectional signal transmitted to all the connected neurons. A NN can be configured to solve specific problems, such as pattern recognition or data classification ones, through a learning process, which implies the adjustment of the information locally stored in each neuron. In this study, we considered the multilayer perceptron network with a feed-forward completely connected three-layer architecture. The training was performed by using the back-propagation algorithm, which is one the most popular training method, successfully applied to a large variety of real-world classification tasks.

4.2. The Experimental Protocol

To illustrate the experimental protocol followed in all the experiments related to the use of feature ranking techniques, let us consider the first DNA microarray dataset, namely *Breast*, and the ranking provided by the first univariate measure, namely CS: by using this feature ranking technique, we generated different representations for such dataset, each containing an increasing number of features. More specifically, we generated 10 representations in the following way: in the first one, the samples were represented by using the first n_1 features in the ranking, in the second one by using the first n_2 features, in the third one the first n_3 features and so on. The considered feature numbers are the following: 5, 10, 50, 100, 200, 500, 1000, 2000, 5000, 10,000. The same procedure has been repeated for the other univariate measures taken into account. The whole process has been applied to the other DNA microarray datasets considered in this study. Summarizing, for each microarray dataset, we produced 5 different feature rankings, each used to generate 10 additional representations, totaling 50 different representations. Each of them has been used in the experiments for evaluating the obtainable classification results, applying the previously described classification schemes.

As for the experimental protocol relative to the use of the heuristic search algorithms, both *Sequential Forward Floating Search* and *Fast Correlation-Based Filter* produced an additional representation of each microarray dataset, in which the samples were represented by using only the features respectively selected. Thus, such algorithms allowed us to generate 2 additional representation of each dataset. On the contrary, *Minimum Redundancy Maximum Relevance* method requires setting a-priori the number N of feature to be selected and then it finds the best feature subset with such a cardinality. Thus, to effectively compare the results of this method with those relative to the other ones, we decided to consider for N the same set of values used in the experiments for the ranking-based feature-selection approaches. Moreover, for the sake of comparison, we have also considered, for each microarray dataset, the values of N respectively provided by the other two heuristic search algorithms.

4.3. Experimental Findings

For the sake of conciseness, we reported in the tables only the best results obtained in each experiment. Please note that in all the tables we indicated with RR and NF the recognition rate and the number of features, respectively, and we put in bold the best result obtained for each dataset. Tables 2 and 3 show the results obtained by using the decision tree classifier: the first table refers to ranking-based feature-selection techniques, while the second one to heuristic search algorithms. For comparison purposes, the last column of both tables indicates the recognition rate relative to the use of the whole feature set. Similarly, Tables 4 and 5 show the results obtained by using the random forest classifier, Tables 6 and 7 those obtained by using the K-Nearest Neighbor classifier, while Tables 8 and 9 report the results relative to the Neural Network classifier. Note that the Neural Network classifier has not always been able to complete the training phase when all the features have been used. In fact, in the case of Breast, Leukemia and Lung datasets, the back-propagation algorithm did not produce significant results due to the enormous complexity of the feature space. This situation has been evidenced in Tables 8 and 9 inserting an asterisk instead of the result.

The analysis of the results reported in the tables suggests the following considerations:

- there is not a single classification scheme or a single feature-selection method that outperforms all the others but, for each microarray dataset, the effectiveness of the whole classification system is due to the combined effect of both feature-selection method and classification scheme;
- for each classification scheme, the best results obtained by using ranking methods are generally comparable with those relative to the use of heuristic search, even if there are in many cases significant differences among the results of the ranking methods as well as among those of the heuristic search ones;
- the number of selected features is generally much less than the total number of available features, while the classification results are always significantly better than those corresponding to the use of the whole feature set: this confirms that for the analysis of DNA microarray datasets, the application of features selection techniques not only reduce the complexity of the feature space, but also allow users to significantly improve classification performance;
- accepting slight variations in the recognition rate (less than 2%), it is possible to obtain very significant reductions in the number of selected features and, therefore, in the computational cost of the classification system. Consider, for instance, the Ovarian dataset for which the best result, equal to 96.85% with 500 features, was obtained by using the feature ranking method based on GR measure with a neural classifier. In the same experimental condition, but selecting only 200 features, the recognition rate was 96.01%. Thus, a reduction of 60% in the dimension of the feature space resulted in a reduction of less than 1% in the recognition rate. We obtained similar results in most of the experiments performed in this study. See, for instance, the results on the Brest dataset with RF classifier and ranking-based feature-selection methods (Table 4): the best results, equal to 89.70% with 200 features, was obtained with CS method, while GR method provided a very similar recognition rate, equal to 89.44% but using only 50 features. Please note that many of these results are not shown in the tables since, for each set of experiments, we reported only the one providing the best recognition rate.

Finally, Table 10 summarizes the best result obtained for each microarray dataset by applying the feature-selection techniques considered in this study. For comparison purposes, the table also shows the best results obtained by using the whole feature set. It is interesting to note that in all the experiments the random forest classifier provided the best results without feature selection: this outcome is in good accordance with the theory, since it is known that such classifier performs very well when the feature number is very high. The data in Table 10 also show that the best performance with feature selection was provided by KNN and NN classifiers. It should be noted, however, that the best results provided by the other two classification schemes, namely DT and RF, are slightly lower, but obtained by selecting in the average a smaller number of features.

Table 2. Best results of ranking-based feature-selection methods with DT classifier.

Dataset	CS		GR		IG		SU		Re		All Features
	RR	NF	RR	NF	RR	NF	RR	NF	RR	NF	RR
Breast	87.48	10	**87.90**	10	87.10	5	87.52	5	87.34	5	81.4
Colon	88.23	200	**88.23**	50	87.74	200	88.23	100	81.94	5	80.02
Leukemia	**98.54**	80	89.03	5	87.50	5	88.75	5	93.06	5	81.9
Lung	97.29	5	96.80	5	96.91	10	96.91	10	**98.62**	5	93.7
Lymphoma	83.65	500	84.43	500	84.74	500	**86.04**	200	82.55	500	78.4
Ovarian	90.74	5	89.49	10	**91.99**	5	88.84	5	88.31	200	85.8

Table 3. Best results of heuristic search algorithms for feature selection with DT classifier.

Dataset	FCBF		MRMR		SFFS		All Features
	RR	NF	RR	NF	RR	NF	RR
Breast	86.95	215	**88.16**	5	84.09	163	81.4
Colon	89.59	14	81.53	100	**89.60**	18	80.02
Leukemia	**84.93**	51	83.47	200	84.44	51	81.9
Lung	96.49	128	**96.99**	10	95.39	163	93.7
Lymphoma	78.43	319	72.86	319	**78.75**	308	78.4
Ovarian	88.42	18	85.05	200	**89.56**	20	85.8

Table 4. Best results of ranking-based feature-selection methods with RF classifier.

Dataset	CS		GR		IG		SU		Re		All Features
	RR	NF	RR	NF	RR	NF	RR	NF	RR	NF	RR
Breast	**89.70**	200	89.44	50	89.56	200	89.60	200	89.56	200	84.69
Colon	86.94	5	**88.87**	50	85.48	10	87.26	5	85.48	10	82.25
Leukemia	98.19	500	98.06	500	97.92	500	**98.26**	1000	97.92	500	90.69
Lung	**99.45**	50	99.45	100	99.45	100	99.45	100	99.45	100	98.07
Lymphoma	90.47	200	91.30	100	**92.29**	10	91.77	100	**92.29**	10	81.92
Ovarian	92.69	10	92.87	50	**92.92**	5	92.18	5	**92.92**	5	87.91

Table 5. Best results of heuristic search algorithms for feature selection with RF classifier.

Dataset	FCBF		MRMR		SFFS		All Features
	RR	NF	RR	NF	RR	NF	RR
Breast	**91.56**	215	90.35	50	90.79	163	84.69
Colon	88.15	14	**85.89**	50	88.87	18	82.25
Leukemia	98.40	51	86.32	200	**98.61**	51	90.69
Lung	**99.45**	128	99.45	500	99.45	163	98.07
Lymphoma	90.78	319	80.78	319	**91.51**	308	81.92
Ovarian	**94.58**	18	88.91	200	93.40	20	87.91

Table 6. Best results of ranking-based feature-selection methods with kNN classifier.

Dataset	CS		GR		IG		SU		Re		All Features
	RR	NF	RR	NF	RR	NF	RR	NF	RR	NF	RR
Breast	90.21	5	**91.96**	50	90.21	100	90.21	50	89.86	200	81.11
Colon	87.10	5	**91.94**	10	85.48	5	88.71	500	87.10	50	75.80
Leukemia	98.61	200	98.61	500	**98.61**	100	98.61	200	97.22	200	86.11
Lung	99.45	2000	98.90	10	**99.45**	5	98.90	50	98.90	5	92.22
Lymphoma	**88.54**	100	85.42	100	88.54	500	88.54	500	87.50	500	78.75
Ovarian	91.67	200	**92.59**	200	90.28	200	90.74	2190	92.59	500	88.89

Table 7. Best results of heuristic search algorithms for feature selection with kNN classifier.

Dataset	FCBF		MRMR		SFFS		All Features
	RR	NF	RR	NF	RR	NF	RR
Breast	90.75	215	**90.75**	10	79.37	133	81.11
Colon	85.4	14	**86.45**	100	85.08	18	75.80
Leukemia	**98.74**	51	85.14	200	97.71	51	86.11
Lung	99.86	128	99.17	163	**100**	163	92.22
Lymphoma	**89.68**	319	67.24	128	89.53	308	78.75
Ovarian	**90.92**	18	87.92	200	89.56	20	88.89

Table 8. Best results of ranking-based feature-selection methods with NN classifier.

Dataset	CS		GR		IG		SU		Re		All Features
	RR	NF	RR	NF	RR	NF	RR	NF	RR	NF	RR
Breast	89.51	500	89.23	1000	**90.21**	100	89.09	500	89.58	2000	*
Colon	82.26	5	**86.13**	10	82.58	5	81.94	50	84.19	5	73.87
Leukemia	**98.61**	500	98.61	1000	98.61	1000	98.61	1000	**98.61**	500	*
Lung	99.45	50	99.45	200	99.45	50	99.45	100	**100**	500	*
Lymphoma	**96.88**	200	95.42	500	95.83	200	95.83	500	95.21	500	28.12
Ovarian	95.19	1000	96.85	500	95.65	1000	95.83	1000	**97.69**	1000	87.13

Table 9. Best results of heuristic search algorithms for feature selection with NN classifier.

Dataset	FCBF		MRMR		SFFS		All Features
	RR	NF	RR	NF	RR	NF	RR
Breast	**90.56**	215	89.69	163	90.24	163	*
Colon	**86.45**	14	83.95	50	84.19	18	73.87
Leukemia	**99.44**	51	95.97	200	97.29	51	*
Lung	**100**	128	**100**	128	100	163	*
Lymphoma	97.92	319	92.45	319	**97.92**	308	28.12
Ovarian	92.41	18	**92.89**	200	90.51	20	87.13

Table 10. Summary of the best results.

Dataset	Without Feature Selection			With Feature Selection			
	RR	NF	cls	RR	NF	cls	f-sel
Breast	84.69	17,816	RF	91.96	50	KNN	GR
Colon	82.25	2000	RF	91.94	10	KNN	GR
Leukemia	90.69	7129	RF	99.44	51	NN	FCBF
Lung	98.07	12,533	RF	100	128	NN	FCBF
Lymphoma	81.92	4026	RF	97.92	308	NN	SFFS
Ovarian	87.91	2190	RF	96.85	500	NN	GR

5. Conclusions and Future Works

In this study, we performed a wide experimental comparison on the combined effect of both feature-selection and classification techniques applied to different DNA microarray datasets, paying particular attention to evaluate how much the number of selected features can affect the obtainable classification performance.

Since DNA microarray experiments typically generate a very large number of features (in the order of thousands) for a limited number of patients (in the order of hundreds or less), the classification task is very complex and typically require the application of a feature-selection process to reduce the complexity of the feature space and to identify a subset of distinctive features.

In our experiments, we considered six microarray datasets, namely *Breast*, *Colon*, *Leukemia*, *Lymphoma*, *Lung* and *Ovarian*, and four classification schemes, namely decision trees, random forest, nearest neighbor, and multilayer perceptron. Such datasets represent an effective test bed, since they exhibit a large variability as regards the number of attributes, the number of classes (two or multiple classes problems) and the number of samples. The considered classifiers provide a broad overview of the results obtainable with the most efficient and widely used classification schemes.

As for the feature-selection process, taking into account the complexity of this problem in the case of thousands of features, we focused our analysis on standard feature ranking techniques, which evaluate each feature singularly, estimating intrinsic properties of the data. These techniques typically provide general solutions, performing well on a wide family of classifiers, even if they could discard relevant features that are weakly correlated with the target class, but highly informative when

combined with other features. The performance of the classification methods was evaluated by using feature sets of increasing size, obtained by selecting the features according to their position in the ranking. Despite this simplified choice, we obtained very interesting results even in comparison with those produced by the other three state-of-the-art feature-selection methods considered in this study, namely the *Sequential Forward Floating Search*, the *Fast Correlation-Based Filter* and the *Minimum Redundancy Maximum Relevance*.

The results confirmed that the feature-selection process plays a key role in this application field and that a reduced feature set allowed us to significantly improve classification performance. They also showed that by accepting a slight reduction of the classification rate, it is possible to further reduce the feature set, significantly improving the computational efficiency of the whole classification system.

As a future work, we would like to improve the classification system by using better performing strategies, such as those based on classifier ensembles [28–30]. The use of Learning Vector Quantization networks could also provide interesting results, as they are generally considered a powerful pattern recognition tool and give the advantage of providing an explicit representation of the prototypes of each class [31] (in this case, a prototype represents the specific gene expression values characterizing tissue and cell samples).

Author Contributions: Conceptualization, N.D.C., C.D.S., F.F. and A.S.d.F.; Methodology, N.D.C., C.D.S., F.F. and A.S.d.F.; Software, S.R., A.S.d.F.; Validation, N.D.C., C.D.S., F.F., S.R. and A.S.d.F.; Formal Analysis, N.D.C., C.D.S., F.F. and A.S.d.F.; Investigation, N.D.C., C.D.S., F.F. and A.S.d.F.; Data Curation, S.R., A.S.d.F.; Writing—Original Draft, C.D.S.; Supervision, C.D.S.; Project Administration, C.D.S. and F.F.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Seijo-Pardo, B.; Bolón-Canedo, V.; Alonso-Betanzos, A. Using a feature selection ensemble on DNA microarray datasets. In Proceedings of the 24th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2016), Bruges, Belgium, 27–29 April 2016; pp. 277–282.
2. Bolón-Canedo, V.; Sánchez-Maroño, N.; Alonso-Betanzos, A.; Benítez, J.; Herrera, F. A review of microarray datasets and applied feature selection methods. *Inf. Sci.* **2014**, *282*, 111–135. [CrossRef]
3. Bolón-Canedo, V.; Morán-Fernández, L.; Alonso-Betanzos, A. An insight on complexity measures and classification in microarray data. In Proceedings of the 2015 International Joint Conference on Neural Networks (IJCNN), Killarney, Ireland, 12–16 July 2015; pp. 1–8.
4. Piatetsky-Shapiro, G.; Tamayo, P. Microarray Data Mining: Facing the Challenges. *SIGKDD Explor. Newsl.* **2003**, *5*, 1–5. [CrossRef]
5. Saeys, Y.; Inza, I.; Larrañaga, P. A review of feature selection techniques in bioinformatics. *Bioinformatics* **2007**, *23*, 2507–2517. [CrossRef] [PubMed]
6. Golub, T.R.; Slonim, D.K.; Tamayo, P.; Huard, C.; Gaasenbeek, M.; Mesirov, J.P.; Coller, H.; Loh, M.L.; Downing, J.R.; Caligiuri, M.A.; et al. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science* **1999**, *286*, 531–537. [CrossRef] [PubMed]
7. Statnikov, A.; Wang, L.; Aliferis, C.F. A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. *BMC Bioinform.* **2008**, *9*, 319–328. [CrossRef] [PubMed]
8. Cilia, N.D.; De Stefano, C.; Fontanella, F.; Scotto di Freca, A. A ranking-based feature selection approach for handwritten character recognition. *Pattern Recognit. Lett.* **2018**. [CrossRef]
9. Wang, Y.; Klijn, J.G.; Zhang, Y.; Sieuwerts, A.M.; Look, M.P.; Yang, F.; Talantov, D.; Timmermans, M.; van Gelder, M.E.M.; Yu, J.; et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet* **2005**, *365*, 671–679. [CrossRef]
10. Alon, U.; Barkai, N.; Notterman, D.A.; Gish, K.; Ybarra, S.; Mack, D.; Levine, A.J. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci. USA* **1999**, *96*, 6745–6750. [CrossRef] [PubMed]

11. Alizadeh, A.A.; Eisen, M.B.; Davis, R.E.; Ma, C.; Lossos, I.S.; Rosenwald, A.; Boldrick, J.C.; Sabet, H.; Tran, T.; Yu, X.; et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **2000**, *403*, 503–511. [CrossRef] [PubMed]
12. Eisen, M.B.; Brown, P.O. DNA arrays for analysis of gene expression. *Methods Enzymol.* **1999**, *303*, 179–205. [PubMed]
13. 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.* **2002**, *62*, 4963–4967. [PubMed]
14. Petricoin, E.F.; Ardekani, A.M.; Hitt, B.A.; Levine, P.J.; Fusaro, V.A.; Steinberg, S.M.; Mills, G.B.; Simone, C.; Fishman, D.A.; Kohn, E.C.; et al. Use of proteomic patterns in serum to identify ovarian cancer. *Lancet* **2002**, *359*, 572–577. [CrossRef]
15. De Stefano, C.; Fontanella, F.; Marrocco, C.; Scotto di Freca, A. A GA-based feature selection approach with an application to handwritten character recognition. *Pattern Recognit. Lett.* **2014**, *35*, 130–141. [CrossRef]
16. Chandrashekar, G.; Sahin, F. A survey on feature selection methods. *Comput. Electr. Eng.* **2014**, *40*, 16–28. [CrossRef]
17. Miao, J.; Niu, L. A Survey on Feature Selection. *Procedia Comput. Sci.* **2016**, *91*, 919–926. [CrossRef]
18. Liu, H.; Setiono, R. Chi2: Feature Selection and Discretization of Numeric Attributes. In Proceedings of the Seventh International Conference on Tools with Artificial Intelligence (ICTAI), Herndon, VA, USA, 5–8 November 1995; IEEE Computer Society: Washington, DC, USA, 1995; pp. 388–391.
19. Kononenko, I. Estimating Attributes: Analysis and Extensions of RELIEF. In Proceedings of the European Conference on Machine Learning, Catania, Italy, 6–8 April 1994; pp. 171–182.
20. Hall, M. Correlation-based Feature Selection for Machine Learning. Ph.D. Thesis, University of Waikato, Hamilton, New Zealand, 1999.
21. Quinlan, J.R. Induction of Decision Trees. *Mach. Learn.* **1986**, *1*, 81–106. [CrossRef]
22. Gutlein, M.; Frank, E.; Hall, M.; Karwath, A. Large scale attribute selection using wrappers. In Proceedings of the IEEE Symposium on Computational Intelligence and Data Mining (CIDM 2009), Nashville, TN, USA, 30 March–2 April 2009; pp. 332–339.
23. Yu, L.; Liu, H. Feature Selection for High-dimensional Data: A Fast Correlation-based Filter Solution. In Proceedings of the Twentieth International Conference on International Conference on Machine Learning, Washington, DC, USA, 21–24 August 2003; pp. 856–863.
24. Peng, H.; Long, F.; Ding, C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *27*, 1226–1238. [CrossRef] [PubMed]
25. Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I.H. The WEKA data mining software: An update. *SIGKDD Explor.* **2009**, *11*, 10–18. [CrossRef]
26. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
27. Gini, C. Measurement of Inequality of Incomes. *Econ. J.* **1921**, *31*, 124–126. [CrossRef]
28. De Stefano, C.; Folino, G.; Fontanella, F.; Scotto di Freca, A. Using Bayesian networks for selecting classifiers in GP ensembles. *Inf. Sci.* **2014**, *258*, 200–216. [CrossRef]
29. De Stefano, C.; D'Elia, C.; Scotto di Freca, A.; Marcelli, A. Classifier Combination by Bayesian Networks for Handwriting Recognition. *Int. J. Pattern Recognit. Artif. Intell.* **2009**, *23*, 887–905. [CrossRef]
30. De Stefano, C.; Fontanella, F.; Scotto di Freca, A. A Novel Naive Bayes Voting Strategy for Combining Classifiers. In Proceedings of the 2012 International Conference on Frontiers in Handwriting Recognition, Bari, Italy, 18–22 September 2012; pp. 467–472. [CrossRef]
31. De Stefano, C.; D'Elia, G.; Marcelli, A. A dynamic approach to learning vector quantization. In Proceedings of the 17th International Conference on Pattern Recognition, Cambridge, UK, 23–26 August 2004; Volume 4, pp. 601–604.