Expert Systems with Applications 39 (2012) 10780–10795

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Comparing the dimensionality reduction methods in gene expression databases

Helyane Bronoski Borges [a,b], Júlio Cesar Nievola [b,*]

[a] UTFPR – Universidade Tecnológica Federal do Paraná, Brazil
[b] PPGIa – Pontifícia Universidade Católica do Paraná (PUCPR), Brazil

ARTICLE INFO

Keywords:
Attribute selection
Random projection
Gene expression database

ABSTRACT

Dimensionality reduction has been applied in the most different areas, among which the data analysis of gene expression obtained with the microarray approach. The data involved in this problem is challenging for machine learning algorithms due to a small number of samples and a high number of attributes. This paper proposes a preprocessing phase by means of attribute selection and random projection method in microarray data. Experimental results are promising and show that the use of these methods improves the performance of classification algorithms.

1. Introduction

The available technologies to process and analyze the information have been allowing people to collect and store information of varied domains. The storage of great amounts of data became possible, but the analysis techniques for understanding the data and visualization of these collections of data was not able to develop in the same proportion (Fayyad, Haussler, & Stolorz, 1996).

Intending to facilitate the analysis and visualization of data, as well as the discovery of useful knowledge for decision making purposes is that tools like data mining emerged, in other words, computation tools that seek for knowledge hidden in databases. This knowledge is said not to be trivial, because they would not be found or noticed by systems of simple analysis, and it is of character ignored to his/her miner.

Data mining is one of the stages of a larger process, named Knowledge Discovery in Databases – KDD. This process is constituted by several stages which can be divided into three main parts: know, pre-processing, data mining and post-processing, that together allow for the knowledge discovery.

According to the common sense, the use of a larger amount of information is more advantageous since, ideally, it would be possible to get a more refined learning. However, this situation may not be true, for computational limitations or for deficiency in the data gathering or it cannot be even advantageous in cases where some data are redundant or irrelevant. Unfortunately, in a large part of the cases it is not possible to know, in advance, which the necessary subset of the data is important for obtaining the best result (Dash & Liu, 1997).

The subject of the amount of data to be used by a learning method can be interpreted as a dimensionality issue, considering two possible aspects to be treated: the amount of instances and the set of attributes. Those aspects form one of the characteristics of gene expression database, obtained from the microarray of DNA technique, where data is formed by a very big number of attributes (genes) and a small number of instances (samples).

To deal with this problem, data dimensionality reduction is used, in order to minimize the volume of data to be treated, regarding the number of attributes, and to increase the generalization capability of the learning methods with the elimination of irrelevant and/or redundant data.

Several techniques of data dimensionality reduction have been proposed and evaluated, mainly in the context of supervised inductive learning (Kohavi & John, 1998; Witten, Ian, & Frank, 2005).

To reduce the dimensionality, there are two basic approaches: transformation of attributes and selection of attributes. Algorithms for attributes transformation create new attributes starting from transformations or combinations of the original group of attributes. Selection algorithms, as their own name suggests, select according to a certain criterion the best subset of the original group of attributes.

Frequently, attributes transformation precedes the selection, then an algorithm for attribute selection eliminates the irrelevant attributes more according to a certain criterion, reducing the dimensionality.

The choice between selection and transformation of attributes depends on the domain of application and on the available training data.

Attributes selection is one of the techniques that has been contributing to an increase in the practical application of methods for Learning of Machine (Liu, Motoda, & Yu, 2003).

Moreover, the application of the attributes selection in gene expression database can still increase the comprehensibility of

* Corresponding author. Tel.: +55 41 3271 1669; fax: +55 41 3271 2121.
E-mail addresses: helyane@utfpr.edu.br (H.B. Borges), nievola@ppgia.pucpr.br (J.C. Nievola).

0957-4174/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2012.03.015

the generated results, identifying the influence of each selected attribute.

This paper provides a comparative study of attribute selection methods applied in five gene expression database with similar characteristics. Two main approaches are used for the attribute selection: the filter approach and the wrapper approach. Those approaches differ in the way the attributes' subsets are tested. The filter approach removes attributes in agreement with the characteristics of the data while in the wrapper approach a learning algorithm to test the candidate subset.

2. Dimensionality reduction approach

2.1. Attribute selection

Attribute selection is one of the most important data pre-preprocessing techniques, being often used. It reduces the number of attributes to be used, removes irrelevant, redundant and noisy data, and brings immediate effects to the application at hand: it improves data mining algorithm's speed, improves precision and comprehensibility of the results.

Through the attribute selection a subset of M attributes out of the N original attributes is chosen, such that $M \leqslant N$, in a way the characteristics space is reduced according to a pre-established condition (Liu et al., 2003). Attribute selection tries to guarantee that data arriving to the mining step are of good quality. A typical attribute selection process is based on four basic steps, as shown in Fig. 1.

The subset generation is a search procedure that creates candidates attributes subsets based on a search strategy to be evaluated. Each generated subset is evaluated and compared with the previous best one, according to an evaluation criterion. If the new subset is better than the old one, it is replaced. If the new subset is worse than the old one, it is discarded and the old subset remains as the best choice. The generation and evaluation process is repeated until some stopping criterion is reached. When it happens, the best subset found needs to be validated through some a priori knowledge or different test samples by a real or synthetic dataset (Liu & Motoda, 1998; Liu et al., 2003; Liu & Yu, 2005).

This process's nature is dependent on two basic aspects. First, one should decide the search starting point (or points), which will determine the search direction. The search can start with an empty set and steadily adding attributes to the existing set (forward search), or it can start with a full set of attributes and removing one attribute at each iteration (backward search). Another possibility is to start with a predefined set of attributes and adding/removing attributes at each step. When the search starts with a randomly selected subset of attributes, it can avoid the trap of a local minimum. The second decision to be made is regarding the search strategy. For a dataset with N attributes, there are 2^N possible subsets. This search space grows exponentially, making it prohibitive, even for a moderate value of N. According to this line of reasoning the search algorithms can be divided in three main groups: exponential, sequential and random algorithms.

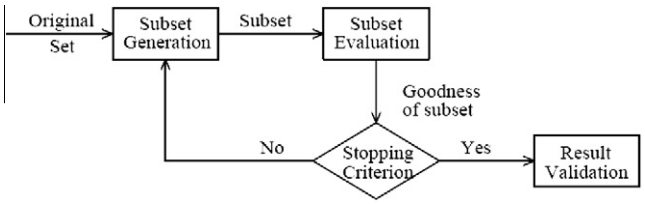

Fig. 1. Attribute selection steps (Kohavi & John, 1998).

Exponential algorithms, such as the exhaustive search, use all the possible combinations of the attributes before returning the resulting attribute subset. They are computationally infeasible, due to the rapidly (exponential) growth of the running time as the number of available attributes arises (Liu & Yu, 2005).

Sequential algorithms, such as the forward sequential selection and the backward sequential selection, are very efficient in solving many problems of attribute selection. Their main disadvantage is that they don't take into account attribute interaction.

Forward sequential selection starts the quest for the best subset attribute with an empty set of attributes. Initially, attribute subsets with only one attribute are evaluated, and the best attribute A^* is selected. This attribute A^* is then combined with all other attributes (in pairs), and the best attribute subset is selected. The search goes on, always adding one attribute at a time to the best attribute subset already selected, until no improvement on the quality of the attribute subset is possible.

Backward sequential selection, contrarily to forward sequential selection, starts the search for the subset of optimal attributes with a set encompassing all attributes, and at each iteration one attribute is removed from the actual solution, until no further improvement on the quality of the solution can be made.

Each new created subset needs to be evaluated using the evaluation criterion. The quality of a subset can be computed according to a certain criterion (for instance, a selected optimum subset through the use of one criterion could not be optimal according to another criterion). In a general way, the evaluation criteria can be categorized in two groups, based on whether it is dependent or not on the mining algorithm that will be applied to the resulting optimal attribute subset.

The filter approach (Fig. 2) characterizes the independent criterion (Kohavi & John, 1998). It tries to evaluate an attribute (or a subset of attributes) exploring intrinsic characteristics of the training data, without any commitment to the mining algorithm.

The most used independent criterions are: distance measures, information measures, dependency measures and consistency measures.

The distance measure is also known as separability, divergence, or discrimination measure. For a two-class problem, an attribute X is preferred to another attribute Y if X induces a greater difference between the two-class conditional probabilities than Y; if the difference is zero, then X and Y are indistinguishable. An example is the Euclidean distance measure (Kohavi & John, 1998; Liu et al., 2003; Liu & Yu, 2005).

The information measure determines the information gain from an attribute. The information gain from an attribute X is defined as the difference between the prior uncertainty and expected posterior uncertainty using X. Attribute X is preferred to attribute Y if the information gain from attribute X is greater than that from attribute Y (e.g., entropy measure) (Kohavi & John, 1998).

The dependency measure is also known as correlation or similarity measure. This measure quantifies how much two attributes are associated, implying that the knowledge of one attribute allows the prediction of the second attribute. In attribute selection context, the best evaluated attribute is the one that best predicts the class (the value of the goal attribute). It measures the ability to predict the value of one variable given the value of another one. For classification purposes, the best attribute is the one that better predicts the class. An attribute X is better than another attribute Y if

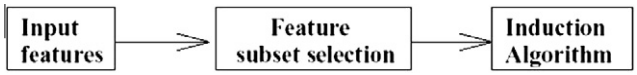

Fig. 2. The filter approach to attribute subset selection (Kohavi & John, 1998).

the association between attribute X and class C is higher than the association between Y and C.

One of the algorithms that use this measure is CFS (Correlation-based Feature Selection) (Kohavi & John, 1998). This algorithm evaluates the importance of an attribute subset based on the individual predictive ability of each attribute and the correlation degree among them. The merit equation in this case is given by Eq. (1):

$$Merit_s = \frac{k\bar{r}_{cf}}{\sqrt{k + k(k-1)\bar{r}_{ff}}} \quad (1)$$

In Eq. (1), *Merit* is the heuristic of the attribute subset S containing k attributes, $\bar{r}_{cf}$ is the average attribute-class correlation ($f \in S$), and $\bar{r}_{ff}$ is the average attribute-attribute inter-correlation (Kohavi & John, 1998).

The difference between CFS and other filter algorithms is that while the general filter algorithms supply an independent score for each attribute, CFS presents a heuristic reward of the attribute subset and informs the best subset found.

The consistency measure is characteristically different from the above measures because of its heavy reliance on the class information and the use of the Min-Features bias in selecting a subset of features (Liu & Yu, 2005). These measures attempt to find a minimum number of features that separates classes as consistently as the full set of features can. An inconsistency is defined as two instances having the same feature values but different class labels. As example, an inconsistency in X' and S is defined by two instances in S being equal when considering only the features in X' and belonging to different classes. The aim is, thus, to find the minimum subset of features leading to zero inconsistencies (Liu et al., 2003). The inconsistency count of an instance is defined as in Eq. (2):

$$IC_{x'}(A) = X'(A) - \max_k X'_k(A) \quad (2)$$

In Eq. (2), $X'(A)$ is the number of instances in S equal to A using only the features in X' and $X'_k(A)$ is the number of instances in S of the class k equal to A using only the features in X'. The inconsistency rate of a feature subset in a sample S is then given by Eq. (3):

$$IR(X') = \frac{\sum_{AES} IC_{X'}(A)}{|S|} \quad (3)$$

This is a monotonic measure in the sense: $X_1 \subset X_2 \Rightarrow IR(X_1) \geqslant IR(X_2)$. A possible evaluation measure is then presented in Eq. (4):

$$J(X') = \frac{1}{IR(X') + 1} \quad (4)$$

This measure is in the range [0,1] and can be evaluated in $O(|S|)$ time using a hash table.

The dependent criterion, characterizes the wrapper approach shown in Fig. 3 (Lin & Gunopulos, 2003). It requires a predefined mining algorithm within the attribute selection and uses its performance when applied in the selected subset to evaluate the quality of the selected attributes.

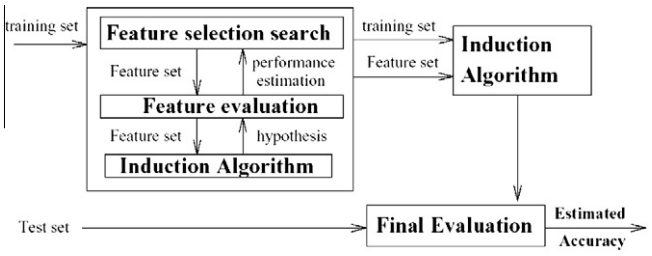

Fig. 3. The wrapper approach to attribute subset selection (Lin & Gunopulos, 2003).

The stopping criterion establishes when the attribute selection process should be stopped. It can be done when the search is over, or when the goal is reached, where the goal can be an specific situation (maximum number of characteristics or maximum number of iterations), or when a sufficiently good subset is found (for instance, a subset could be sufficiently good if the classification error rate is under some threshold for a given task).

2.2. Random projection

The idea of the random projection method is simple: given a matrix X, data dimensionality can be reduced by projecting it through the origin onto a lower-dimensional subspace, formed by a set of random vectors (Kohavi & John, 1998): $A_{[n\times k]} = X_{[n\times m]} * X_{[m\times k]}$, where k represents the amount of columns of the reduced matrix.

The random projection method is motivated by the Johnson–Lindenstrauss Lemma (Lin & Gunopulos, 2003):

For any $0 < \varepsilon < 1$ and any integer n, let k be a positive integer such that: $k \geqslant 4(\varepsilon^2/2 - \varepsilon^3/3)^{-k}\ln n$. Then for any set W of n points in R^m there is a map $f: R^m \to R^k$ such that for all u, $v \in W$, $(1-\varepsilon)\|u-v\|^2 \leqslant \|f(u) - f(v)\|^2 \leqslant (1+\varepsilon)\|u-v\|^2$.

Analyzing the theorem, the equation allows deducing that a set of n points in high-dimensional Euclidean space can be mapped down onto an $O(\log n/\varepsilon^2)$ dimensional subspace such that the distances between the points are approximately preserved (Liu & Motoda, 1998).

The elements in R are Gaussian-distributed, where a Gaussian distribution is defined as $G(u,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$, where μ is the average and σ is the standard deviation of the distribution.

Achlioptas (2001) has proposed two simpler distributions (Achlioptas, 2001):

$$r_{i,j} = \begin{cases} +1 \text{ with prob. } 1/2 \\ -1 \text{ with prob. } 1/2 \end{cases} \quad (5)$$

and

$$r_{i,j} = \sqrt{3} * \begin{cases} +1 \text{ with prob. } 1/6 \\ 0 \text{ with prob. } 2/3 \\ -1 \text{ with prob. } 1/6 \end{cases} \quad (6)$$

These simple distributions reduce computational time for the calculation of $X * R$ (Lin & Gunopulos, 2003). With this method, the original m-dimensional data are projected in subset $k (k \ll d)$ (Bertoni & Valentini, 2005). In this way, the original matrix $X_{n\times m}$ is projected onto the matrix $R_{m\times k}$ obtaining the reduced matrix $A_{n\times k}$.

2.3. Related works

Borges and Nievola (2005), Borges (2006) using the same database used by Alizadeh and colleagues in their experiments, applied attribute selection algorithms looking for improvement in the results of the classification algorithms. To apply the attribute selection methods in that work, the authors divided the work in two phases: the search for attributes subsets and the evaluation of the found subsets. Two methods of search were used: the sequential search and the random search. For subsets evaluation it was used two main approaches: the filter and the wrapper approaches. As evaluation measures of the filter approach the dependence and the consistency measure were used. The wrapper approach used the data mining algorithm itself for the subsets evaluation, where the algorithms Naïve Bayes, Bayesian Networks, C4.5, Decision Table and k-NN (for the $k = 1, k = 3, k = 5$ and $k = 7$) were chosen. These experiments it was noticed a great variation in the number of attributes selected according to the search and evaluation

method of the subset under evaluation. Comparing the obtained results, the attribute selection improved the results of the classification in practically all of the situations. The use of the wrapper evaluation method produced best results in a constant way. Such result was mainly evident with the application of the algorithm wrapper used during the attribute selection.

Bertoni and Valentini (2005) applied the dimensionality reduction method to aid in the clustering analysis clustering of gene expression data (Bertoni & Valentini, 2005). The idea in applying that method consisted of evaluating the stability of the clustering discovered when applied with all data and when applying the random projection method. The method was applied in the DLBCL-tumor database and the results showed that the application of the method can be used to identify state stable data clustering, without a priori knowledge and without suppositions about the data distribution, independent of the clustering algorithm. Besides, the authors state, through the experiments, that the use of the method can be useful in the identification of a probable number of clustering and to help the biomedical researches.

In another work Borges and Nievola (2007), applied the same attribute selection algorithms in the DLBCL-Tumor and BLBCL-Outcome database. In these experiments, it has been observed that the use of attribute selection, despite the huge number of attributes and small number of samples, has lead to a better performance of the classifiers in almost all cases and schemes on both datasets. Random search is not advisable, considering that its results are hard predictable. The use of sequential search, on the other hand, consistently achieved improvements in the classification. The results also indicate that the sequential forward search algorithm is to be used, considering not only its best computational results, but also that it matches biological knowledge saying that in general few genes interact in order to give raise to a characteristic, in this case the cancer.

Chuang et al. (2008) using binary particle swarm optimization (IBPSO) to implement attribute selection, and the *K*-nearest neighbor (*K*-NN) method serves as an evaluator of the IBPSO for gene expression data classification problems in 11 datasets. In yours experiments the results show that the method used reduces the number of attribute. The When applied the classification was obtained highest classification accuracy in 9 datasets.

Reynes et al. (2008) was proposed the new method to attribute selection. This method is based on SELDI-TOF mass spectrometry. As there are a huge number of possibilities a genetic algorithm is used to select to the best population. The experiments are executed in tow datasets and gives promising results.

Borges and Nievola (2009) propose a preprocessing phase by means of random projection method in microarray data. The dimension of the new attributes' group was defined of random form for the five bases. For the formation of this new group two criterions were chosen: one using a fixed number of attributes and other using a desired percentage of attributes. These results are statistically analyzed through the use of hypotheses test with pairs samples. To compare the results of the use of random projection against the results obtained by the algorithms classification applied on the databases using all of the attributes, the arithmetic average was calculated.

Zhu et al. (2010) propose a model-based approach to estimate the entropy of class variables on the model. For this used multivariate normal distributions to fit the data, because multivariate normal distributions have maximum entropy among all real-valued distributions with a specified mean and standard deviation and are widely used to approximate various distributions. Given that the data follow a multivariate normal distribution, since the conditional distribution of class variables given the selected features is a normal distribution, its entropy can be computed with the log-determinant of its covariance matrix. Because of the large number of genes, the computation of all possible log-determinants is not efficient. The authors propose several algorithms to largely reduce the computational cost. The results experiments on seven gene data sets show the accuracy of the multivariate Gaussian generative model and the efficiency of algorithms. A small increase in the performance of the results of the classifiers compared with the results obtained when using the database with all the attributes was noticed. The algorithm SVM was the one that presented best results statistically significant and in none of the experiments it was considered worse statistically than the algorithm base Naïve Bayes.

3. Method

The overall process can be seen in Fig. 4. This process is divided into three steps. The first step focus on the databases classification using all attributes. The second step is the attribute selection and classification using the selected attributes (Borges & Nievola, 2008; Hall, 2000). In the third step random projection is used and then the chosen attributes are used for classification (Hall, 2000). Each step of this process is detailed described in the following sections.

3.1. Description databases

Five database were selected, four of them regarding the study of the lymphoma and one about leukemia. The data used in the experiment was extracted from the Kent Ridge Biomedical Data Set Repository (http://sdmc.i2r.a-star.edu.sg/rp/).

In Fig. 5 it is shown some characteristics of the used bases, indicating the number of attributes, amount of samples, number of classes and the size of each subsample, according to the target class. The bases are identified by the following names: DLBCL (Alizadeh et al., 2000), DLBCL-Tumor (Shipp et al., 2002), DLBCL-Outcome (Shipp et al., 2002), DLBCL-NIH (Rosenwald et al., 2002) and ALL/AML (Golub et al., 1999).

The DLBCL database consists of gene expression data, obtained from the microarray approach, in which the authors studied Diffuse Large B Cell Lymphoma (DLBCL). This is the most common subtype of non-Hodgkin lymphoma more common for a group of cancers (malignancies). Two distinct forms of DLBCL cells were identified, which had gene expression patterns indicated by two different stages of B cell. One type of gene expression characterizes the germinal center B cell and the other type of gene expression is the activation of B cells [ALI00]. The data set consists of 47 examples, with 24 of them belonging to the germinal center B cell, while 23 belong to the B cell activation Each example is described by 4026 genes, all with a numeric value, besides the target attribute (Alizadeh et al., 2000).

The DLBCL-Tumor database consists of two types of lymphoma: Diffuse Large B Cell Lymphoma (DLBCL) and Follicular Lymphoma (FL). The set has 7129 attributes and 77 samples from patients, 58 belonging to the group of DLBCL and 19 belonging to the group LF (Shipp et al., 2002).

The DLBCL-Outcome database is composed by 58 samples, 32 belonging to the cured patient's group and 26 to the patient's group no cured, using 7129 attributes. This database presents the result of the DLBCL patient's treatment after 5 years. The data in this base identify which patients had success in the treatment and were cured and the ones with no cure (Shipp et al., 2002).

The DLBCL-NIH database is formed by 240 patient samples (102 alive and 138 not alive) and 7399 attributes (genes). The objective with this database is to analyze the patient's survival condition with Diffuse Lymphoma of Large Cell B after the chemotherapy (Rosenwald et al., 2002).

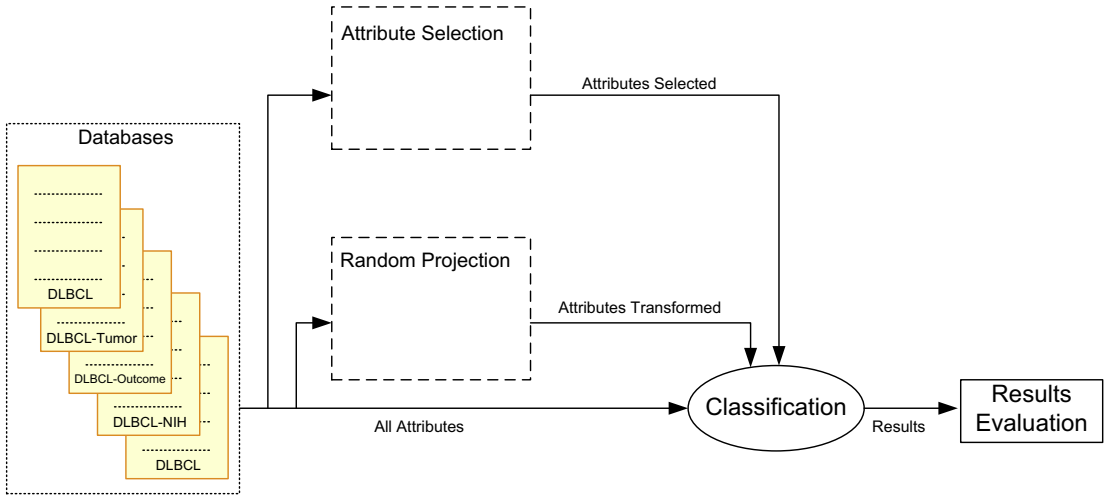

Fig. 4. Method of the experiments.

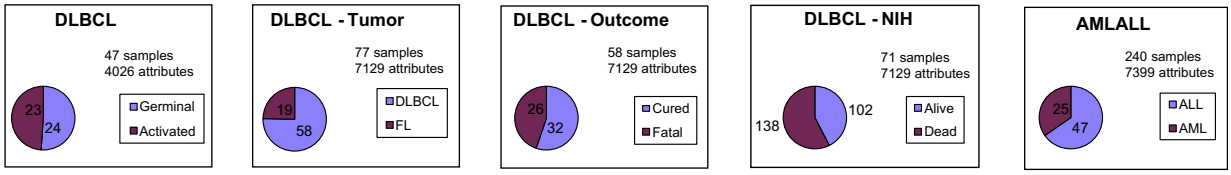

Fig. 5. Representation of the division of the database class.

The ALL/AML database analyzes two types of acute leukemia (Golub et al., 1999). The set has 7129 attributes (genes) and 72 samples divided into 47 belonging to the ALL leukemia (acute lymphoblastic leukemia) and 25 belong to AML (acute myeloid leukemia).

3.2. Pre-processing

At this stage, the data dimensionality reduction was made using two different methods: the attributes selection and random projection.

3.2.1. Attribute selection

For attribute selection two kinds of search were used: sequential (S) and random (R). Used evaluation measures were based on filter and wrapper approach (W). For the filter approach dependency (D) and consistency (C) measures were chosen. For the wrapper approach, as it uses the mining algorithm for the subset evaluation, four classification algorithms were used: Naïve Bayes (NB), induction decision tree executed by the algorithm C4.5, SVM and *k*-NN. All those algorithms were run considering their parameters configured with the standard values (as they can be found in Weka environment). For the SVM classifier the parameters used were: with value same to 1.0 and epsilon with value 1.0^2-12. For the *k*-NN classifier $k = 1$, $k = 3$, $k = 5$ and $k = 7$ values were used.

The sequential search applied in the experiments used a forward greedy search beginning with any attribute. In this paper the strategy adopted was to stop the process when the subsequent addition (or deletion) of any attribute doesn't provide an improvement in the quality measure.

The random search used in these experiments uses a simple genetic algorithm (Goldberg, 1989). The evaluation solution found (fitness function or aptitude function) is based on the number of instances correctly classified. The initial population was created based on a uniform random distribution. The group of the parameters was established a priori, with the size of the population and the number of generations being defined as 20, and the crossover and mutation probabilities defined as 0.6 and 0.033, respectively.

Fig. 6 shows how the attribute selection was made combining the search methods and evaluation measures of subsets generated.

3.2.2. Random projection

Fig. 7 shows the steps to implement the random projection method. The dimension of the new attributes' group was defined in a random way for the five databases. For the formation of this new group two criteria were chosen: one using a fixed number of attributes and the other using a desired percentage of attributes. For the number of attributes the following values were chosen: 10, 15, 30, 45, and 71 attributes. As desired percentage of attributes the following values were chosen: 3%, 10%, 20%, 25%, and 50%. In this way it is possible to compare the results among the bases through the percentage and also of a fixed number of attributes independently of the base size. According to this, 10 new groups were generated.

For the same number of attributes, the method was executed 10 times varying the seed of the random matrix generation. Ten random values were chosen that varied from 5 to 65. The distribution used for the calculation of the random matrix is described in the Eq. (6). Making a combination between the number of attributes and the number of the random seed one can see that 100 new subsets were generated for each base.

3.3. Data mining

The objective of the paper is the attribute reduction in microarray databases using the attribute selection and the random projection. However, it is important to remember that, if the dimensionality reduction is excessive, the classifier can have its discrimination power reduced. Therefore, it is important to compare

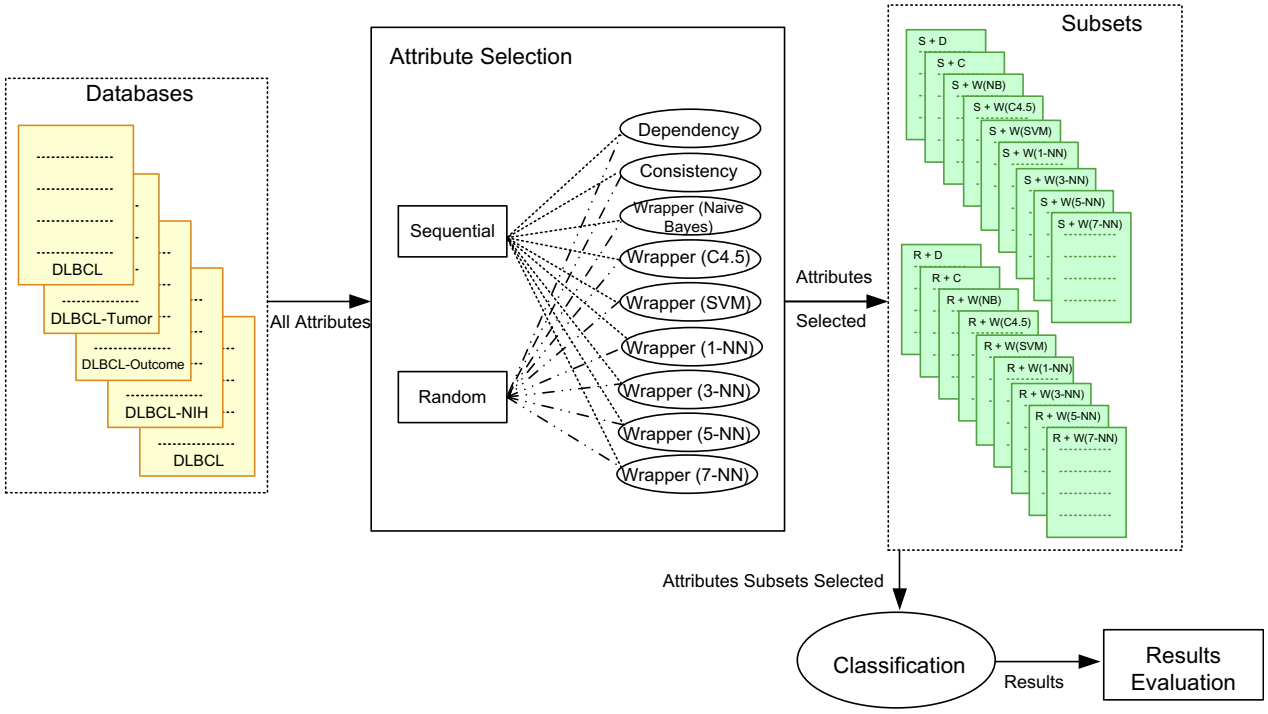

Fig. 6. Steps for execution the attributes selection.

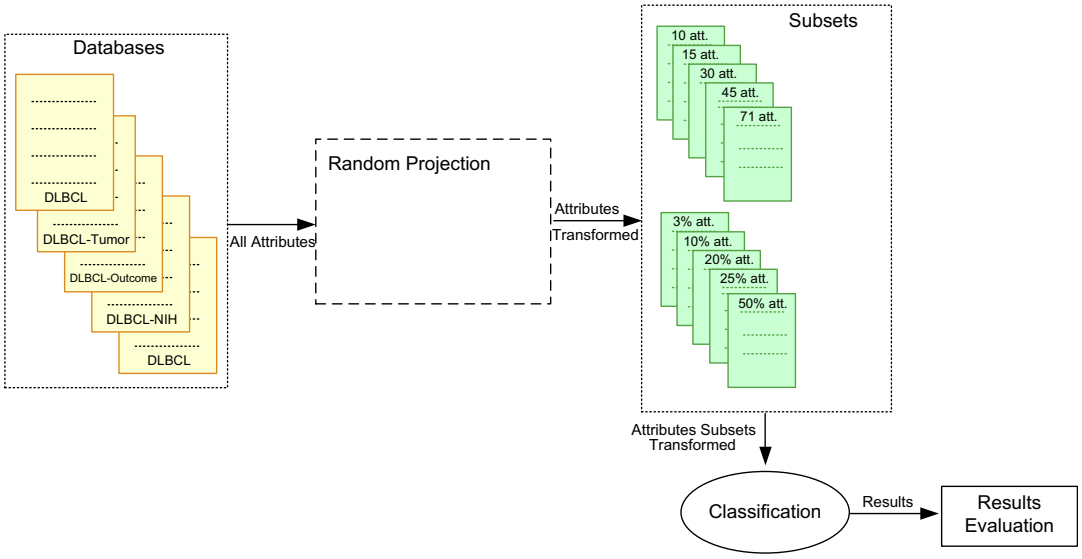

Fig. 7. Steps for execution the random projection.

the variation of the classifier behavior regarding the database with all attributes, so that it is possible to estimate the ideal dimensionality of the dataset, for a defined classifier based on these results.

The five databases were submitted to 4 classifiers: Naïve Bayes, C4.5, SVM and k-NN (for $k = 1$, $k = 3$, $k = 5$, and $k = 7$) as a way to compare the classification rate of the classifier using all attributes in relation to its behavior when using the subsets generated by the reduction dimensionality methods. The indicated algorithms were chosen because they belong to different learning paradigms.

3.4. Post-processing

The results were evaluated through estimate of average success for each classifier using 10-fold stratified cross-validation. Besides, for all the results, statistical analyses were used to evaluate the significance of the results through the test of hypothesis with p-value 0.05, which establishes the misclassification threshold and, at the same time, defines the null hypothesis rejection area.

The term statistics significance used in this work refers to the result of an algorithms comparison using the paired t-test statistics, where the fundamental objective is to evaluate the behavior of the differences observed in each element (Zhu et al., 2010).

4. Results and discussion

The objective here is to compare the results of the different combinations of search methods and evaluation criteria of the gen-

erated subsets through the acting of on some classifiers in the five databases.

The results were obtained by running the algorithms on the original database and the attributes' subsets generated by the methods of dimensionality reduction (attributes selection and random projection).

The five chosen databases were submitted to the four classifiers. In the tables containing the results, the note"*" indicates that a result is significantly statistically worse than the result of the standard algorithm (Naïve Bayes) and the result in bold indicates that the result is significantly statistically better compared to the standard algorithm.

4.1. Results of classification using all attributes

Initially the classification algorithms were executed on the original data for comparison purposes with the attribute selection algorithms. Table 1 shows the results when applying the classifiers on the data using all attributes in databases.

It is observed that the algorithms 1-NN, 3-NN and 7-NN, in the attributes subsets of the DLBCL database had results statically worse than those obtained with the algorithm Naïve Bayes. For the attributes' subsets of the DLBCL-Tumor database the algorithm SVM presented the best results statistically compared to the base algorithm.

Analyzing the results of all classifiers on DLBCL-Outcome and DLBCL-NIH databases, one shows that all results are lower when compared with the other three ones. This low classification may have happened due to the type of data being used. In these databases the result is based only on the interaction of some genes. One could argue that the evolution of the treatment is also dependent on some other factors like the type of treatment, specific characteristics of the disease, the patient's characteristics (as age, sex) etc. These information, however, were not available in the database and, therefore, were not used. In the attributes subsets generated for these databases the results of classifiers were also low; however those results are considered statistically equivalent.

Analyzing the statistical results on the attributes subsets of the ALL/AML the algorithms C4.5, 3-NN, 5-NN and 7-NN had the worst results.

4.2. Results of the attribute selection

Two approaches were used for the attribute selection: filter and wrapper. The filter approach was applied with dependency and consistency evaluation measures. For the wrapper approach the mining algorithms evaluation measures selected were: Naïve Bayes (W (NB)), C4.5 (W (C4.5)), SVM (W (SVM)) and the *k*-NN (W (1-NN)), (W (3-NN)), (W (5-NN)) and (W (7-NN)). As search criteria the sequential search (S) and the random search (R) were used.

It is possible to observe in the Fig. 7 the great variation in the number of attributes selected by each method. The sequential search method caused a great reduction in the number of attribute selected compared to the random search method. This is an algorithm characteristic of the algorithm, since it adds an attribute at each step and then tends to find a small attributes subset. For the random search it used a genetic algorithm which had a great number of attributes selected. That happens due the random choice of the attributes, that does not privilege a specific subset size.

4.2.1. Filter approach

For each subset of attributes the classification algorithms were applied. Table 2 shows the success rate of the classifiers in the subsets of attributes of the DLBCL database using the filter approach. The results show that the sequential search method and dependency evaluation measure and consistency evaluation measure were similar compared to the Naïve Bayes algorithm (used as the standard in the experiments). For the attributes subset generated by random search and the dependency evaluation measure, two results were considered statistically worse: the 1-NN and 7-NN. However, with the consistency evaluation measure results were significantly worse in the algorithms 1-NN and 3-NN. If they are compared to the original database results, these three classification algorithms presented the worst results in the database.

Table 3 shows the results of the attributes subsets of the DLBCL-Tumor database where most of the results ara statically similar compared to the Naïve Bayes algorithm. Only two results are considered worse, C4.5 in the method that used the sequential search and dependency evaluation measure, and when SVM was used when the sequential search and consistence evaluation measured.

In Table 4 the results of the use of filter approach on each attribute subset of the DLBCL-Outcome database are presented. Only two of those results are considered worse, when compared to the Naïve Bayes algorithm: the results obtained with the C4.5 algorithm and with the algorithm 7-NN. Both results were worse when combined with the sequential search method and the dependency evaluation measure.

In Table 5 it is observed that the algorithm *k*-NN presented worse results to the subset of attributes of the base of data

Table 1
Results of the classification of the databases with all of the attributes (precision, in %).

Databases	Algorithms						
	Naïve Bayes	C4.5	SVM	1-NN	3-NN	5-NN	7-NN
DLBCL	97.50 ± 7.91	77.00 ± 23.71	98.00 ± 6.32	75.50 ± 21.27*	77.00 ± 17.51*	75.00 ± 23.69	73.00 ± 18.74*
DLBCL-Tumor	80.54 ± 10.70	72.50 ± 16.15	**96.07 ± 6.34**	84.11 ± 13.56	93.21 ± 9.85	89.82 ± 9.93	91.07 ± 8.54
DLBCL-Outcome	42.00 ± 24.81	53.33 ± 11.55	54.33 ± 20.73	45.67 ± 24.65	38.67 ± 24.61	47.67 ± 23.83	53.00 ± 24.47
DLBCL-NIH	59.58 ± 11.96	52.08 ± 9.87	63.75 ± 11.46	51.25 ± 10.22	49.17 ± 11.59	50.83 ± 9.78	50.00 ± 7.61
ALL/AML	98.57 ± 4.52	78.93 ± 15.63*	98.57 ± 4.52	84.64 ± 18.14	83.39 ± 12.88*	83.39 ± 12.88*	77.86 ± 11.30*

Table 2
Classification results of the attributes subsets of the DLBCL database after the attribute selection using the filter approach (precision, in %).

Subsets	Algorithms						
	Naïve Bayes	C4.5	SVM	1-NN	3-NN	5-NN	7-NN
S + D	100.00 ± 0.00	76.50 ± 30.00	100.00 ± 0.00	98.00 ± 6.32	98.00 ± 6.32	96.00 ± 8.43	96.00 ± 8.43
S + C	94.00 ± 9.66	89.00 ± 11.74	89.50 ± 11.17	93.50 ± 10.55	96.00 ± 8.43	93.50 ± 10.55	93.50 ± 10.55
R + D	97.50 ± 7.91	80.00 ± 16.83	98.00 ± 6.32	78.00 ± 15.31*	85.50 ± 13.83	78.50 ± 22.37	73.00 ± 18.74*
R + C	89.00 ± 15.06	83.00 ± 13.17	85.00 ± 14.14	72.50 ± 8.90*	66.00 ± 16.96*	75.00 ± 15.63	72.50 ± 19.90

Table 3
Classification results of the attributes subsets of the DLBCL-Tumor database after the attribute selection using the filter approach (precision, in %).

Subsets	Algorithms						
	Naïve Bayes	C4.5	SVM	1-NN	3-NN	5-NN	7-NN
S + D	96.07 ± 3.34	86.79 ± 12.26*	97.50 ± 5.27	93.57 ± 9.00	95.89 ± 6.63	97.32 ± 5.66	97.32 ± 5.66
S + C	93.57 ± 6.80	90.89 ± 10.96	76.61 ± 5.48*	91.96 ± 14.80	92.14 ± 9.00	89.64 ± 9.99	90.89 ± 8.64
R + D	81.96 ± 11.86	80.36 ± 13.86	96.07 ± 6.34	86.79 ± 10.75	91.79 ± 11.88	91.07 ± 8.54	90.89 ± 8.64
R + C	80.54 ± 13.56	79.46 ± 11.73	94.82 ± 6.71	85.36 ± 10.23	90.36 ± 13.45	90.89 ± 10.96	89.46 ± 10.56

Table 4
Classification results of the attributes subsets of the DLBCL-Outcome database after the attribute selection using the filter approach (precision, in %).

Subsets	Algorithms						
	Naïve Bayes	C4.5	SVM	1-NN	3-NN	5-NN	7-NN
S + D	84.00 ± 13.59	68.33 ± 17.23*	77.33 ± 14.38	70.00 ± 28.50	73.00 ± 27.60	73.67 ± 20.45	73.67 ± 17.17*
S + C	70.33 ± 25.07	68.33 ± 15.34	70.33 ± 20.58	77.00 ± 17.10	75.67 ± 12.28	65.67 ± 15.48	69.00 ± 13.15
R + D	35.67 ± 26.11	46.33 ± 20.21	59.33 ± 22.54	44.00 ± 27.25	40.33 ± 28.26	46.00 ± 24.23	56.33 ± 27.86
R + C	44.00 ± 23.03	63.67 ± 20.69	57.67 ± 21.14	44.00 ± 27.25	47.67 ± 23.83	49.33 ± 21.93	51.33 ± 22.67

Table 5
Classification results of the attributes subsets of the DLBCL-NIH database after the attribute selection using the filter approach (precision, in %).

Subsets	Algorithms						
	Naïve Bayes	C4.5	SVM	1-NN	3-NN	5-NN	7-NN
S + D	72.92 ± 10.62	66.25 ± 8.88	65.42 ± 8.80	55.83 ± 8.15*	57.08 ± 8.80	57.08 ± 7.62*	57.50 ± 7.03*
S + C	66.25 ± 8.21	66.67 ± 6.80	61.67 ± 11.08	53.75 ± 11.19*	57.08 ± 6.53*	58.75 ± 6.35	57.08 ± 6.23
R + D	61.67 ± 11.42	57.50 ± 9.98	56.25 ± 7.67	59.17 ± 7.81	49.58 ± 9.10	50.83 ± 10.90	49.58 ± 9.31*
R + C	60.42 ± 9.05	50.42 ± 10.66	66.25 ± 12.02	53.75 ± 8.88	47.92 ± 11.83*	52.50 ± 11.98	52.50 ± 6.27

DLBCL-NIH. The algorithm k-NN with $k = 1$ presented worse results in two methods of selection of attributes, both by sequential search with the two measures of evaluation of the approach filter. The algorithm 3-NN presented worse results in the method of sequential search with the evaluation measure and in the random search with the consistence measure of evaluation. In algorithm 5-NN the worst results happens with the method of sequential search and the measure of evaluation dependency and in the method of random search with the measure of evaluation dependency. Finally, the algorithm 7-NN presented worse results in the method of sequential and random search, but just in the measure of evaluation dependency.

Table 6 displays the results obtained in the experiments accomplished with the ALL/AML database. In that table almost all the results of the subsets of attributes are statically equivalent, being only three of them considered worse when compared with the algorithm Naïve Bayes: the algorithm C4.5 in the method of sequential search with the measure of evaluation dependency, the algorithm 7-NN in the method of random search with the measure of evaluation dependency and the algorithm SVM in the method of sequential search with the measure of evaluation consistence.

Calculating the average result of all of the executions, in Table 7, it is possible to verify that when the subset where the sequential search and the dependency measure were used just the algorithm SVM had values statically equivalent to the algorithm Naïve Bayes. Instead, in the subset R + C that algorithm had precision statically worse when compared with the algorithm base Naïve Bayes. In the subsets R + D and R + C the results of the algorithm SVM were statically better than the result of the other algorithms.

Analyzing those results, it is possible to observe also that the method of sequential search and the measure of evaluation dependency presented better results. That result is easier to be seen when one compares the average of the results for all executions of that approach (according to Fig. 8).

4.2.2. Wrapper approach

The results using the approach wrapper on the subsets of data are showed in Tables 8–12. These results were obtained by the method of sequential search and method of random search in each one of the databases.

Table 8 presents the results obtained in the DLBCL database where one can observe that the experiments with sequential search using the algorithm C4.5 got worse results when compared to the algorithms Naïve Bayes and 7-NN. When using random search the SVM algorithm was the only one that had result statically equivalent to the base algorithm base Naïve Bayes, the remaining of the results were statically worse.

Table 9 shows the results obtained in the DLBCL-Tumor database. With sequential search only the result of the algorithm

Table 6
Classification results of the attributes subsets of the ALL/AML database after the attribute selection using the filter approach (precision, in %).

Subsets	Algorithms						
	Naïve Bayes	C4.5	SVM	1-NN	3-NN	5-NN	7-NN
S + D	98.57 ± 4.52	81.96 ± 11.56*	98.57 ± 4.52	94.46 ± 9.83	95.71 ± 9.64	94.46 ± 9.83	94.46 ± 9.83
S + C	95.71 ± 9.64	93.04 ± 9.98	81.96 ± 11.56*	93.04 ± 12.04	94.64 ± 11.33	93.04 ± 12.04	93.04 ± 9.98
R + D	95.89 ± 6.63	83.04 ± 11.33	97.14 ± 6.02	83.39 ± 19.27	83.39 ± 12.88	84.64 ± 14.27	81.79 ± 11.93*
R + C	80.54 ± 13.69	89.11 ± 10.66	88.75 ± 13.10	78.04 ± 18.06	80.89 ± 10.72	83.75 ± 13.48	75.54 ± 12.43

Table 7
Average precision with selection of attributes. using the approach filter (precision, in %).

Subsets	Algorithms						
	Naïve Bayes	C4.5	SVM	1-NN	3-NN	5-NN	7-NN
S + D	90.3 ± 11.60	75.9 ± 8.72*	87.8 ± 15.57	82.4 ± 18.52*	83.9 ± 18.17*	83.7 ± 17.76*	83.8 ± 17.60*
S + C	84.0 ± 14.38	81.6 ± 12.94	76 ± 10.66	81.9 ± 17.16	83.1 ± 16.67	80.1 ± 16.58	80.7 ± 16.69
R + D	74.5 ± 26.5	69.5 ± 16.53	**81.4 ± 21.56**	70.3 ± 18.16	70.1 ± 23.41	70.2 ± 20.46	70.3 ± 17.23
R + C	70.7 ± 18.21	73.3 ± 15.90	**78.5 ± 15.79**	66.7 ± 17.26	66.6 ± 19.19	70.3 ± 18.60	68.3 ± 16.22

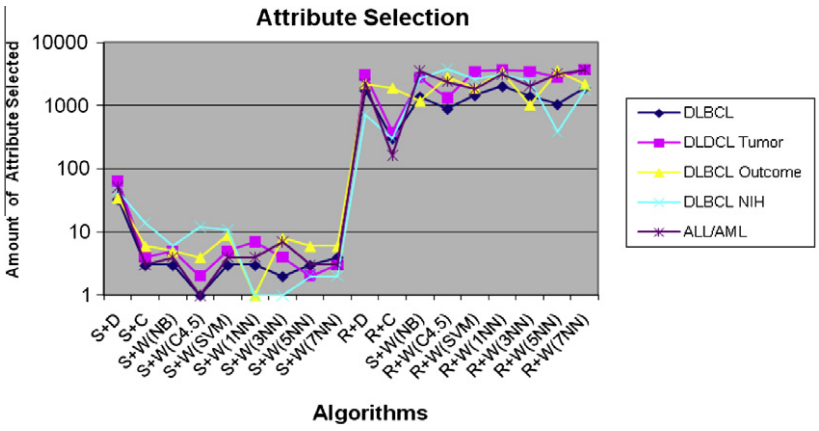

Fig. 8. Number of attributes selected (Obs.: the number of attributes selected is in a logarithmic scale).

Table 8
Classification results of the attributes subsets of the DLBCL database after the attribute selection using the wrapper approach (precision, in %).

Search method	Algorithms						
	Naïve Bayes	C4.5	SVM	1-NN	3-NN	5-NN	7-NN
S	98 ± 6.32	91.5 ± 11.07*	98.0 ± 6.32	98 ± 6.32	98 ± 6.32	98 ± 6.32	100 ± 0
R	100 ± 0	91.5 ± 11.07*	100 ± 0	84.5 ± 14.42*	79.5 ± 16.41*	79 ± 20.25*	81.5 ± 17.65*

Table 9
Classification results of the attributes subsets of the DLBCL-Tumor database after the attribute selection using the wrapper approach (precision, in %).

Search method	Algorithms						
	Naïve Bayes	C4.5	SVM	1-NN	3-NN	5-NN	7-NN
S	100 ± 0	93.4 ± 11.36*	98.75 ± 3.95*	100 ± 0	97.3 ± 5.66*	94.8 ± 6.71*	97.3 ± 5.66*
R	81.9 ± 10.7	89.3 ± 12.57	98.75 ± 3.95	88.04 ± 9.93*	93.2 ± 9.85	92.3 ± 8.87	92.3 ± 8.87

Table 10
Classification results of the attributes subsets of the DLBCL-Outcome database after the attribute selection using the wrapper approach (precision, in %).

Search method	Algorithms						
	Naïve Bayes	C4.5	SVM	1-NN	3-NN	5-NN	7-NN
S	89.67 ± 11.91	90.67 ± 13.68	94 ± 13.5	77.67 ± 11.0*	89.67 ± 11.91	93.33 ± 11.65	86 ± 13.41
R	52.67 ± 21.76	48 ± 19.7	59.67 ± 21.69	52.33 ± 24.95	65.33 ± 19.45	51 ± 19.69	60 ± 21.6

Table 11
Classification results of the attributes subsets of the DLBCL-NIH database after the attribute selection using the wrapper approach (precision, in %).

Search method	Algorithms						
	Naïve Bayes	C4.5	SVM	1-NN	3-NN	5-NN	7-NN
S	75.00 ± 8.33	75.83 ± 7.30	73.3 ± 4.89*	60.42 ± 9.47*	64.17 ± 9.04*	67.08 ± 7.97*	68.33 ± 8.83*
R	64.17 ± 10.24	55.42 ± 7.87*	72.5 ± 6.86	58.75 ± 8.88*	59.58 ± 7.36*	57.92 ± 8.21*	55.42 ± 11.79*

Table 12
Classification results of the attributes subsets of the ALL/AML database after the attribute selection using the wrapper approach (precision, in %).

Search method	Algorithms						
	Naïve Bayes	C4.5	SVM	1-NN	3-NN	5-NN	7-NN
S	100.00 ± 0	94.64 ± 9.11*	97.14 ± 6.02*	100 ± 0	100 ± 0	100 ± 0	100 ± 0
R	100.00 ± 0	93.21 ± 7.18	100 ± 0	90.36 ± 14.68*	87.5 ± 14.19*	87.5 ± 14.19*	86.07 ± 13.49*

Table 13
Average of the executions of the methods of the attributes selection. using the wrapper approach. in the five attributes subsets of the databases (precision, in %).

Search method	Algorithms						
	Naïve Bayes	C4.5	SVM	1-NN	3-NN	5-NN	7-NN
S	93 ± 10.69	89.2 ± 7.64*	92.2 ± 10.74	87.2 ± 17.69*	89.8 ± 14.87	90.6 ± 13.43	90.3 ± 13.59
R	79.7 ± 21.22	75.5 ± 21.91	86.2 ± 18.91	74.8 ± 17.85	77 ± 14.31	73.5 ± 18.23*	75.1 ± 16.37

1-NN was statically equivalent to Naïve Bayes. The results obtained by the other algorithms were statically worse. When using random search results show that the SVM algorithm presented statically better result and the 1-NN the worst result.

In Table 10 the obtained results for DLBCL-Outcome database are presented. In the sequential search (S) method the 1-NN algorithm had statistically worst results compared to the base algorithm. For the random search (R) the 3-NN algorithm stood out with the best evaluation.

Table 11 shows the results obtained with the experiments realized in the DLBCL-NIH database. In the sequential search method the results of the SVM and k-NN, for $k = 1$, $k = 3$, $k = 5$ and $k = 7$ algorithms brought the worst results. By using random search the SVM algorithm had the statistically better result.

Table 12 shows the results obtained in the ALL/AML database. It is observed that C4.5 and SVM algorithms had the worst statistically results with sequential search. On the other side, for random search their results were statically similar to the Naïve Bayes algorithm. Again, the algorithms 1-NN, 3-NN, 5-NN and 7-NN had the statically worst results.

Calculating the average of all of the executions of the experiments (Table 13), it is observed that the use of sequential search with algorithms C4.5 and the 1-NN presented the worst results compared to the base algorithm. If one looks at the random search, algorithm SVM presented the statically the best results and the algorithm 5-NN the worst results.

Comparing the two types of search used, sequential and random, it is observed that the sequential search had better results in the wrapper approach and that is visible when applied to the average of the results (Fig. 9).

The Fig. 10 shows precision rate of the classifiers when used in the original database (with all of the attributes) and in the best and worst cases.

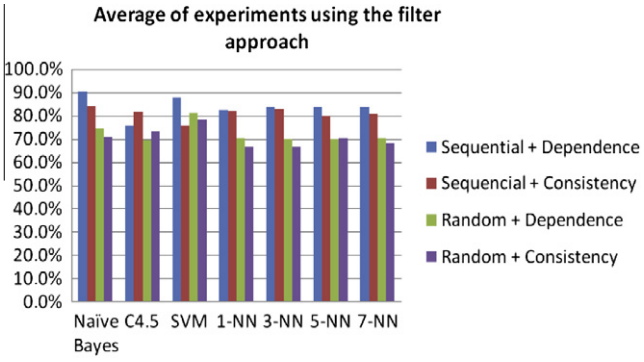

Fig. 9. Average precision using attributes selection with the approach filter.

The use of attribute selection improved the precision rate of the classifier in all of the attributes subsets. The reached precision is clearly superior in the best cases, independent of the classification algorithm that is being used. In the worst cases, the results are very similar when applied to the original database, when all the attributes are used.

4.3. Results of the random projection

The random projection method was applied on five databases as described in Section 4. For each number of attributes chosen, the random projection method was executed ten times to obtain a result. The average of the results of each generated subset was calculated and the results are presented.

The results of the each database are divided in two tables, one of which shows the results obtained when a fixed number of attributes was used for the formation of the new attributes' subset and the other presents the results when a percentage of the attributes was chosen.

Tables 14 and 15 show the results for DLBCL database. Tables 16 and 17 show the results for DLBCL-Tumor database. Tables 18 and 19 show the results for DLBCL-Outcome database. Tables 20 and 21 show the results for DLBCL-NIH database. Finally, Tables 22 and 23 show the results for ALL/AML database.

In the DLBCL, for subsets with 10 attributes chosen, the results of the C4.5, SVM e 7-NN are statically similar compared to the standard algorithm (Naïve Bayes). For the subsets with 15 and 30 attributes the SVM algorithm had better results compared to the Naïve Bayes algorithm and the 1-NN algorithm got worse results. The k-NN algorithm for all k values had statistically lower results compared to the Naïve Bayes algorithm in the experiments which used a subset consisting of 45 attributes and 71 attributes.

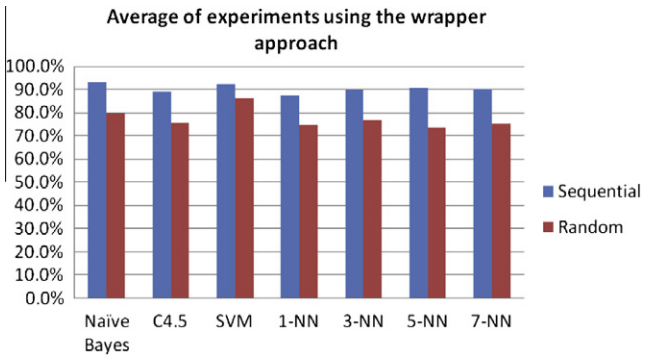

Fig. 10. Average of the executions of the methods of the attributes selection, using the wrapper approach, in the five attributes subsets of the databases (precision, in %).

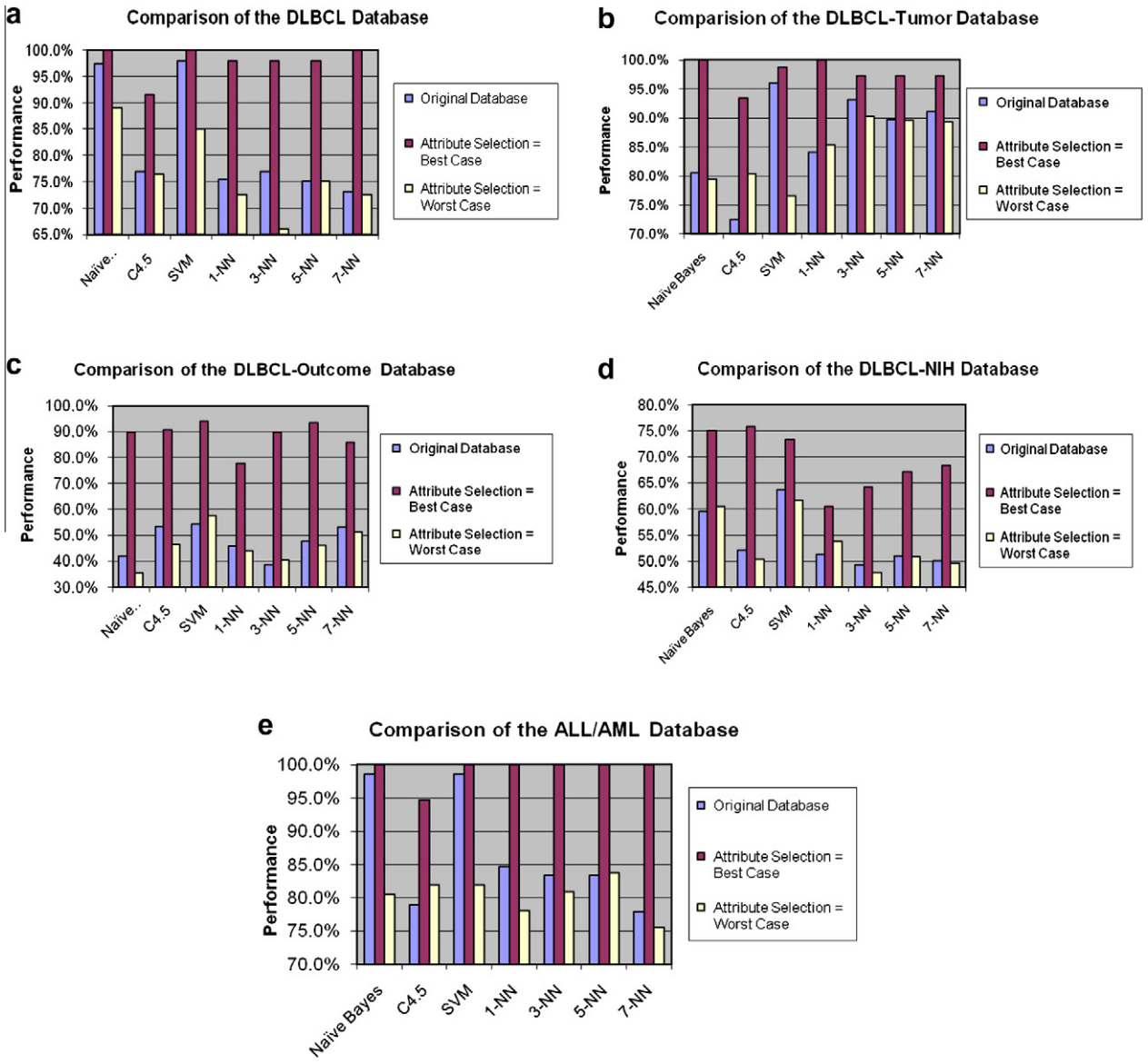

Fig. 11. Comparison of the attributes selection methods in the attributes in the best case and in worst case (precision, in %).

Table 14
Result of the random projection method of the DLBCL database when used a fixed number of attributes for the formation of the attributes subset (precision, in %).

Amount attributes	Algorithms						
	Naïve Bayes	C4.5	SVM	1-NN	3-NN	5-NN	7-NN
10 attributes	73.35 ± 9.92	65.25 ± 8.09	76.40 ± 8.92	66.65 ± 10.59*	66.90 ± 9.79*	68.35 ± 11.36*	69.90 ± 10.13
15 attributes	71.70 ± 10.58	69.35 ± 12.71	**79.70 ± 9.95**	68.70 ± 9.20*	69.80 ± 10.93	74.25 ± 8.87	71.05 ± 11.97
30 attributes	75.25 ± 8.52	70.45 ± 12.73	82.00 ± 8.49	69.95 ± 9.39*	73.20 ± 7.77	71.55 ± 9.69	76.00 ± 11.27
45 attributes	83.40 ± 5.64	76.10 ± 11.54	87.20 ± 4.67	70.15 ± 4.56*	75.15 ± 5.85*	75.50 ± 7.80*	76.75 ± 5.57*
71 attributes	86.25 ± 6.52	75.05 ± 10.92*	86.15 ± 4.66	73.75 ± 7.42*	75.50 ± 8.71*	78.55 ± 7.71*	79.30 ± 8.26*

Table 15
Result of the random projection method of the DLBCL database when using a percentage of attributes for the formation of the attributes' subset (precision, in %).

Amount attributes	Algorithms						
	Naïve Bayes	C4.5	SVM	1-NN	3-NN	5-NN	7-NN
3% attributes	90.35 ± 4.12	79.70 ± 5.08*	90.60 ± 5.61	76.00 ± 5.57*	79.10 ± 4.50*	80.80 ± 6.93*	80.70 ± 9.25*
10% attributes	94.05 ± 2.85	72.60 ± 6.99*	92.95 ± 2.31	81.00 ± 5.33*	86.35 ± 3.50*	86.95 ± 3.00*	84.00 ± 5.88*
20% attributes	94.00 ± 2.22	72.05 ± 7.54*	93.70 ± 3.57	80.70 ± 4.14*	86.55 ± 3.95*	85.85 ± 1.89*	87.80 ± 3.17*
25% attributes	94.35 ± 3.16	73.25 ± 10.59*	93.20 ± 3.24	80.30 ± 3.50*	85.50 ± 2.52*	86.95 ± 2.15*	86.25 ± 3.33*
50% attributes	94.80 ± 1.87	78.90 ± 12.33*	94.35 ± 3.06	82.00 ± 2.25*	85.45 ± 2.19*	87.35 ± 1.72*	87.00 ± 2.48*

Table 16
Result of the random projection method of the DLBCL-Tumor database when using a fixed number of attributes for the formation of the attributes' subset (precision, in %).

Amount attributes	Algorithms						
	Naïve Bayes	C4.5	SVM	1-NN	3-NN	5-NN	7-NN
10 attributes	81.27 ± 5.87	79.56 ± 6.42	78.45 ± 3.79	79.05 ± 8.09	81.86 ± 5.36	83.02 ± 5.25	82.25 ± 6.49
15 attributes	82.20 ± 5.19	79.07 ± 5.42	81.95 ± 5.47	81.36 ± 6.69	82.77 ± 6.18	82.63 ± 5.01	**82.91 ± 6.79**
30 attributes	86.02 ± 5.93	82.28 ± 4.82	87.89 ± 4.47	83.53 ± 5.43	82.41 ± 5.47	84.00 ± 5.64	85.77 ± 4.21
45 attributes	88.36 ± 4.00	82.27 ± 4.60*	**92.61 ± 3.77**	86.77 ± 4.11	85.29 ± 4.44*	87.12 ± 4.72	87.38 ± 4.57
71 attributes	88.59 ± 3.04	83.57 ± 5.14*	**94.62 ± 2.53**	87.55 ± 3.10	85.54 ± 3.75*	88.07 ± 3.98	88.59 ± 4.28

Table 17
Result of the random projection method of the DLBCL-Tumor database when using a percentage of attributes for the formation of the attributes' subset (precision, in %).

Amount attributes	Algorithms						
	Naïve Bayes	C4.5	SVM	1-NN	3-NN	5-NN	7-NN
3% attributes	88.55 ± 1.97	83.36 ± 5.66*	**95.87 ± 1.40**	87.84 ± 2.30	86.87 ± 1.79	88.53 ± 1.78	**90.32 ± 2.23**
10% attributes	88.57 ± 2.31	85.84 ± 6.30	**96.11 ± 1.27**	87.73 ± 2.21	86.28 ± 1.77*	90.37 ± 1.62	89.98 ± 1.93
20% attributes	88.84 ± 1.69	85.21 ± 6.27	**96.79 ± 1.12**	87.86 ± 1.57	86.89 ± 1.51*	90.48 ± 1.72	**91.98 ± 2.33**
25% attributes	87.93 ± 3.10	82.86 ± 6.86*	**96.50 ± 1.29**	87.98 ± 2.10	87.41 ± 2.34	90.12 ± 1.24	**91.52 ± 2.18**
50% attributes	88.77 ± 1.41	86.61 ± 4.46	**96.84 ± 0.99**	88.17 ± 1.38	86.51 ± 0.93*	90.97 ± 1.05	**91.13 ± 1.55**

Table 18
Result of the random projection method of the DLBCL-Outcome database when using a fixed number of attributes for the formation of the attributes' subset (precision, in %).

Amount attributes	Algorithms						
	Naïve Bayes	C4.5	SVM	1-NN	3-NN	5-NN	7-NN
10 attributes	45.50 ± 7.34	**53.00 ± 7.41**	**52.80 ± 6.24**	**53.77 ± 9.40**	**55.07 ± 6.66**	**53.47 ± 6.12**	**52.37 ± 7.41**
15 attributes	45.53 ± 8.11	51.57 ± 5.38	50.63 ± 4.28	**53.90 ± 7.14**	**53.33 ± 4.69**	**52.07 ± 5.57**	**52.47 ± 7.10**
30 attributes	43.37 ± 7.89	47.57 ± 6.28	**48.47 ± 9.77**	**56.30 ± 8.14**	**54.27 ± 5.59**	**51.07 ± 6.58**	**48.33 ± 7.11**
45 attributes	41.77 ± 4.52	43.07 ± 5.34	**47.07 ± 6.30**	**52.27 ± 4.51**	**54.13 ± 6.54**	**52.43 ± 7.07**	**51.43 ± 7.15**
71 attributes	40.50 ± 6.15	**48.97 ± 7.61**	**51.57 ± 5.88**	**52.07 ± 6.50**	**55.00 ± 4.99**	**51.70 ± 7.76**	**51.00 ± 4.83**

Table 19
Result of the random projection method of the DLBCL-Outcome database when using a percentage of attributes for the formation of the attributes' subset (precision, in %).

Amount attributes	Algorithms						
	Naïve Bayes	C4.5	SVM	1-NN	3-NN	5-NN	7-NN
3% attributes	39.73 ± 6.53	45.70 ± 8.59	**52.77 ± 7.15**	**57.10 ± 5.45**	**55.54 ± 4.89**	**55.37 ± 6.99**	**52.23 ± 5.66**
10% attributes	38.80 ± 4.63	**48.37 ± 9.33**	**51.00 ± 2.81**	**54.50 ± 1.71**	**56.57 ± 4.80**	**53.14 ± 3.86**	**49.83 ± 4.54**
20% attributes	39.07 ± 3.56	**48.97 ± 8.46**	**52.00 ± 1.61**	**53.70 ± 2.68**	**57.90 ± 3.04**	**54.24 ± 2.87**	**51.57 ± 2.47**
25% attributes	37.63 ± 4.46	46.83 ± 14.91	**52.30 ± 2.74**	**53.10 ± 2.90**	**57.57 ± 2.34**	**54.47 ± 3.92**	**50.10 ± 3.14**
50% attributes	36.60 ± 2.92	**47.70 ± 8.85**	**52.00 ± 2.25**	**51.63 ± 2.10**	**56.73 ± 2.52**	**52.23 ± 2.73**	**50.50 ± 3.3**

Table 20
Result of the random projection method of the DLBCL-NIH database when used a fixed number of attributes for the formation of the attributes subset (precision, in %).

Amount attributes	Algorithms						
	Naïve Bayes	C4.5	SVM	1-NN	3-NN	5-NN	7-NN
10 attributes	55.63 ± 2.06	56.67 ± 2.17	56.84 ± 0.97	54.17 ± 3.35	54.17 ± 3.29	54.83 ± 3.06	56.34 ± 2.82
15 attributes	56.54 ± 2.77	56.50 ± 2.85	56.54 ± 1.18	54.67 ± 3.93	53.71 ± 2.35*	53.38 ± 2.16*	54.04 ± 3.72
30 attributes	55.63 ± 2.52	54.54 ± 3.29	55.87 ± 4.09	52.58 ± 4.20	54.38 ± 3.30	53.29 ± 3.19	53.12 ± 2.95*
45 attributes	56.67 ± 3.25	54.42 ± 1.77	55.96 ± 4.42	54.50 ± 1.98	54.42 ± 1.67	54.42 ± 2.93*	56.12 ± 2.76
71 attributes	58.54 ± 2.65	52.67 ± 3.46*	56.58 ± 5.67	55.21 ± 2.73*	56.67 ± 3.22	56.38 ± 2.48	56.00 ± 2.76

Table 21
Result of the random projection method of the DLBCL-NIH database when used a percentage of attributes for the formation of the attributes subset (precision, in %).

Amount attributes	Algorithms						
	Naïve Bayes	C4.5	SVM	1-NN	3-NN	5-NN	7-NN
3% attributes	59.83 ± 1.98	55.33 ± 4.61*	59.88 ± 3.43	56.71 ± 2.45*	58.71 ± 1.91	57.21 ± 1.57*	58.79 ± 1.81
10% attributes	59.79 ± 1.76	52.08 ± 2.61*	62.29 ± 3.96	56.50 ± 1.67*	59.54 ± 3.20	59.00 ± 1.79	58.21 ± 2.31
20% attributes	60.50 ± 2.46	54.87 ± 4.31*	62.50 ± 3.90	56.63 ± 1.22*	60.33 ± 2.37	59.00 ± 2.21*	57.87 ± 2.12*
25% attributes	59.87 ± 1.55	54.58 ± 2.48*	**63.38 ± 3.23**	57.46 ± 1.76*	60.33 ± 2.11	58.84 ± 2.65	57.63 ± 2.95*
50% attributes	59.87 ± 1.08	55.46 ± 2.52*	**64.17 ± 2.42**	57.17 ± 1.36*	61.12 ± 2.20	57.75 ± 1.99*	57.75 ± 1.91*

Analyzing Table 15 it is observed that only the SVM algorithm had similar results compared to Naïve Bayes algorithm in all experiments. The results obtained by other algorithms are statistically worse.

If compared to the result of the original database, the random projection method obtained better results in most of the situations. Only when compared to subsets with 10 and 30 attributes the result of the dataset with all attributes was better. Another behavior arises when the random projection was applied using a percentage of attributes, obtaining the best results, both in comparison to a fixed number of attributes as well as to the database with all attributes.

Analyzing the results of the random projection method in the DLBCL-Tumor database (Table 16), the results of the algorithms in the experiments with 10 and 30 attributes are considered similar to the Naïve Bayes algorithm. In the subset consisting of 15 attributes the 7-NN algorithm got statistically better results than other algorithms. In experiments with 45 and 71 attributes SVM algorithm presented statistically better results and algorithms C4.5 and 3-NN the worst results compared to the Naïve Bayes algorithm.

In the experiments that used a percentage of the original attributes to the formation of attributes subsets, the SVM algorithm was statistically better in all cases, compared with the Naïve Bayes algorithm. The 7-NN also had statistically better results in almost all experiments. Besides, it is observed the that 3-NN had the worst results in three experiments, the ones made with 10%, 20% and 50% of the original attributes and the C4.5 algorithm in the attributes subset with 25% of them (Table 17).

Also comparing the original database and the attributes subsets generated by the random projection method, it is observed that in some cases the success rate of the classifier in the original database was better compared with the subsets formed by a fixed number of attributes. Note that in almost all cases that used the random projection method the result of the algorithms were superior.

Analyzing the results of the DLBCL-Outcome database using the random projection method with a fixed number of attributes (Table 18) as well as the percentage of attributes (Table 19), shows that statistically almost all algorithms were considered better than the Naïve Bayes algorithm.

Analyzing the results obtained by the execution of the random projection method in the DLBCL-NIH database when using a fixed number of attributes (Table 20), it is observed that statistically all results obtained in the subset of 10 attributes are statistically equivalent to the Naïve Bayes algorithm. The results obtained in the subset of 15 attributes the 3-NN and 5-NN algorithms are considered significantly worse. The same happens in the subset of 30 attributes with the 7-NN algorithm, the subset of 45 attributes with the 5-NN algorithm and the subset of 71 attributes with the algorithms C4.5 and 1-NN.

When the percentage of attributes was used (Table 21) it is observed that the C4.5 and the 1-NN algorithm had statistically worse results compared to the Naïve Bayes algorithm in all cases. The 5-NN and 7-NN algorithms also had statistically worst in some experiments.

The use of SVM presented statistically better results when applied to subsets with 25% and 50% of attributes.

Analyzing the results obtained by the execution of the random projection method in the ALL/AML database using a fixed number of attributes (Table 22), it is observed that the algorithm C4.5 presented statistically the worst results in all experiments compared to the algorithm Naïve Bayes.

It is noticed that the algorithm 1-NN had worse results in the subsets with 10 and 45 attributes, the algorithm 3-NN in the subset with 71 attributes, the algorithm 5-NN in the subsets with 45 and 71 attributes and the algorithm 7-NN in the subsets with 15, 30, 45 and 71 attributes.

In Table 23 one can observe the results obtained when the percentage of attributes was used. The algorithm C4.5 and the 7-NN had statistically worse results in all experiments. It is also possible to observe that in the subset formed by 3% of attributes the algorithm 1-NN, 3-NN and 5-NN had the results considered statistically worse compared to the base algorithm Naïve Bayes. The same happened in the subset with 20% of attributes with the algorithm 5-NN.

It is observed, although, that the SVM is considered statistically better in all the experiments when a percentage of the original attributes was used for the formation of the subset in the ALL/AML database.

It is observed that in most cases the use of random projection method helped to improve the hit rate, especially when a percentage of attributes was used to form the attributes subset.

From these results, some conclusions can be drawn from experiments that used the random projection method. First, the SVM algorithm showed the best results statistically significant and none of the experiments was considered statistically worse than the Naïve Bayes algorithm.

4.4. General comparison

After an individual analysis of each experiment, the average behavior for all ranking algorithms was evaluated. For the experiments that applied attribute selection results are divided into two approaches: filter and wrapper. Then it is calculated the average overall results for both approaches. For the experiments in which the random projection method was used, the results are divided into two parts: the first for a fixed number of attributes and the second for the percentage of attributes for the formation of new subsets of attributes.

These results are also statistically analyzed using hypothesis tests with paired samples compared on the Naïve Bayes algorithm. All these results are compared with the results obtained by the classification algorithms when applied to the databases with all the original attributes. Table 24 shows the arithmetic mean obtained by the algorithms in the five databases. Analyzing these results it is noted that only the C4.5 algorithm got statistically worse values.

4.4.1. Comparison of the application of the attributes selection

Table 25 shows the average results with attribute selection, when using filter approach, for each classification algorithm in all five datasets.

It is observed that in the datasets DLBCL, DLBCL-NIH and ALL/AML the SVM algorithm had result statically similar to the Naïve Bayes algorithm and other algorithms had results statically worse compared to the base algorithm.

In the DLBCL-Tumor database the algorithms SVM, 3-NN, 5-NN and 7-NN had results statically better and the C4.5 algorithm resulted statically worse, both compared to Naïve Bayes. In the DLBCL-Outcome database just the SVM algorithm achieved better result than Naïve Bayes.

Table 26 shows the average of the results with attribute selection when the wrapper approach was used, for each classification algorithm.

In DLBCL and ALL/AML databases the result was the same as in filter approach. In the DLBCL database only the SVM algorithm had results statically equivalent to the algorithm base, the other algorithms had results statically worse. In the base DLBCL-Tumor the algorithms SVM, 3-NN and 7-NN arrived at results statically better and in the base DLBCL-Outcome dataset all the results were

Table 22
Result of the random projection method of the ALL/AML database when using a fixed number of attributes for the formation of the attributes' subset (precision, in %).

Amount attributes	Algorithms						
	Naïve Bayes	C4.5	SVM	1-NN	3-NN	5-NN	7-NN
10 attributes	80.41 ± 7.56	75.02 ± 7.08*	81.29 ± 7.13	76.71 ± 6.98*	77.98 ± 5.68	78.18 ± 4.88	77.55 ± 5.22
15 attributes	84.09 ± 6.41	75.18 ± 8.48*	84.11 ± 6.68	80.89 ± 4.70	80.68 ± 5.45	80.54 ± 5.47	80.11 ± 4.78*
30 attributes	86.80 ± 5.09	74.38 ± 5.22*	87.68 ± 7.55	85.21 ± 4.06	86.52 ± 2.74	84.77 ± 3.77	82.30 ± 2.97*
45 attributes	89.04 ± 4.94	73.27 ± 6.65*	**93.34 ± 2.88**	85.27 ± 2.63*	87.29 ± 3.29	85.95 ± 2.84*	84.93 ± 2.87*
71 attributes	90.50 ± 2.56	78.59 ± 3.39*	**95.30 ± 3.27**	89.20 ± 3.52	88.84 ± 2.02*	87.12 ± 1.78*	85.73 ± 3.11*

Table 23
Result of the random projection method of the ALL/AML database when using a percentage of attributes for the formation of the attributes' subset (precision, in %).

Amount attributes	Algorithms						
	Naïve Bayes	C4.5	SVM	1-NN	3-NN	5-NN	7-NN
3% attributes	91.45 ± 2.42	80.48 ± 5.97*	**96.30 ± 1.80**	89.29 ± 2.31*	89.47 ± 1.84*	88.77 ± 2.65*	88.07 ± 1.91*
10% attributes	93.00 ± 2.65	83.16 ± 5.40*	**97.14 ± 1.17**	91.88 ± 1.77	91.95 ± 2.31	90.82 ± 1.62	90.06 ± 1.65*
20% attributes	93.43 ± 2.15	81.14 ± 6.62*	**97.86 ± 1.01**	91.86 ± 1.54	93.04 ± 1.00	91.11 ± 1.59*	90.63 ± 1.57*
25% attributes	93.00 ± 2.07	80.84 ± 7.32*	**97.86 ± 0.75**	91.84 ± 1.26	92.90 ± 1.12	91.29 ± 1.88	90.40 ± 1.36*
50% attributes	92.43 ± 1.66	82.34 ± 6.92*	**97.86 ± 0.75**	92.54 ± 1.58	93.15 ± 0.60	92.20 ± 1.21	90.13 ± 1.66*

Table 24
General average for classification of the original databases (precision, in %).

Algorithms						
Naïve Bayes	C4.5	SVM	1-NN	3-NN	5-NN	7-NN
83.0 ± 15.8	78.7 ± 13.3*	85.1 ± 14.2	78.2 ± 17.0*	79.47 ± 16.0	79.1 ± 15.9	79.2 ± 15.5

Table 25
Average precision for algorithm of classification for each database using attribute selection – filer approach (in %).

Databases	Algorithms						
	Naïve Bayes	C4.5	SVM	1-NN	3-NN	5-NN	7-NN
DLBCL	95.1 ± 4.8	82.1 ± 5.3*	93.1 ± 7.1	85.5 ± 12.2*	86.4 ± 14.6*	85.8 ± 10.5*	83.8 ± 12.7*
DLBCL-Tumor	87.7 ± 8.3	84.6 ± 5.1*	91.3 ± 9.8	89.5 ± 4	92.5 ± 2.4	92.2 ± 3.4	92.1 ± 3.5
DLBCL-Outcome	58.5 ± 22.5	61.7 ± 10.5	66.2 ± 9.3	58.8 ± 17.3	59.2 ± 17.8	58.7 ± 13.2	62.6 ± 10.5
DLBCL-NIH	65.3 ± 5.7	60.2 ± 7.8*	62.4 ± 4.6	55.6 ± 2.5*	52.9 ± 4.9*	54.8 ± 3.8*	54.2 ± 3.8*
ALL/AML	92.7 ± 8.2	86.8 ± 5.2*	91.6 ± 7.8	87.2 ± 7.9*	88.7 ± 7.6*	89 ± 5.6*	86.2 ± 9.1*

Table 26
Average precision for algorithm of classification for each database using the attribute selection – wrapper approach (in %).

Databases	Algorithms						
	Naïve Bayes	C4.5	SVM	1-NN	3-NN	5-NN	7-NN
DLBCL	99.0 ± 1.4	91.5 ± 0*	99.0 ± 1.4	91.3 ± 9.5*	88.8 ± 13.1*	88.5 ± 13.4*	90.8 ± 13.1*
DLBCL-Tumor	91.0 ± 12.8	91.4 ± 2.9	98.8 ± 0	94.0 ± 8.5	95.3 ± 3.0	93.6 ± 1.8	94.8 ± 3.5
DLBCL-Outcome	71.2 ± 26.2	69.3 ± 30.2	76.8 ± 24.3	65.0 ± 17.9	77.5 ± 17.2	72.2 ± 29.9	73.0 ± 18.4
DLBCL-NIH	69.6 ± 7.7	65.6 ± 14.4*	72.9 ± 0.6	59.6 ± 1.2*	61.9 ± 3.2*	62.5 ± 6.5*	61.9 ± 9.2*
ALL/AML	100 ± 0	93.9 ± 1.0*	98.6 ± 2.1	95.2 ± 6.8*	93.8 ± 8.8*	93.8 ± 8.8*	93.0 ± 9.8*

statically equivalent. Considering the base DLBCL-NIH, the algorithm SVM had results statically better and the other algorithms were statically worse compared to the Naïve Bayes algorithm.

The average of the results for filter and wrapper approach can be observed in Table 27. It is observed that the algorithms that had the worst results in the two approaches individually also had the worse result overall. The same happens with the best algorithm.

The average classification, for each classification algorithm, in all databases is shown in Table 28. It is observed that the results for C4.5 and 1-NN algorithms are statically worse than the algorithm Naïve Bayes.

Comparing those results with the average of the obtained results of the original database, it is noticed the increase in the success tax. Besides, analyzing and comparing the original results and the results of the attribute selection, it is observed that just the SVM algorithm got results statically equivalent, while the other algorithms resulted in better results with the application of attribute selection.

4.4.2. Comparison of random projection method application

Table 29 shows the results' average of the random projection method when a fixed number of attributes was used for each classification algorithm on the databases.

Through statistical analyses it was possible to observe that the algorithm SVM had better results compared to the algorithm Naïve Bayes in the five databases. It is also observed in the DLBCL-Outcome database that the other algorithms also stood out presenting

better results than the algorithm base. In the other databases the algorithms C4.5 and *k*-NN had worse statistical results.

Table 30 shows the results' average obtained by the random projection method, using a percentage of attributes. In DLBCL database just the algorithm SVM had equivalent results to the algorithm base; other algorithms had results statistically worst. In DLBCL-tumor database the algorithms SVM, 5-NN and 7-NN had better results and the algorithms C4.5, 1-NN and 3-NN worse results. In DLBCL-Outcome database all the results of the algorithms were statistically better than the algorithm Naïve Bayes, but in the DLBCL-NIH and ALL/AML databases only the SVM algorithm SVM better statistical results. For the remaining algorithms the results were statistically worse compared to the base algorithm.

Table 31 shows the general average of the random projection method on each database. It is observed that the statistical results obtained were a lot similar to the results of the two previous tables, which means, in most of the cases the algorithm SVM had better statistical results.

Table 32 presents the general average for the random projection method in all databases. The analysis of these results indicates that all algorithms lead to equivalent results when compared to Naïve Bayes. Comparing these results to the results obtained in the original database one can see that the algorithms Naïve Bayes and SVM presented inferior results. However, analyzing statistically the results obtained of this method with the results obtained of the

Table 27
General average precision of the algorithms of classification for each dataset using attribute selection (in %).

Databases	Algorithms						
	Naïve Bayes	C4.5	SVM	1-NN	3-NN	5-NN	7-NN
DLBCL	97.1 ± 2.8	86.8 ± 6.6*	96.1 ± 4.2	88.5 ± 4.2*	87.6 ± 1.7*	87.2 ± 1.9*	87.3 ± 4.9*
DLBCL-Tumor	89.3 ± 2.3	88.0 ± 4.8	95.0 ± 5.3	91.8 ± 3.2	93.9 ± 1.9	92.9 ± 1.0	93.5 ± 1.9
DLBCL-Outcome	64.8 ± 9.0	65.5 ± 5.4	71.5 ± 7.5	61.9 ± 4.4*	68.4 ± 12.9	65.4 ± 9.5	67.8 ± 7.4
DLBCL-NIH	67.4 ± 3.0	62.9 ± 3.8*	67.7 ± 7.4	57.6 ± 2.8*	57.4 ± 6.3*	58.7 ± 5.4*	58.0 ± 5.4*
ALL/AML	96.4 ± 5.2	90.4 ± 5.0*	95.1 ± 4.9	91.2 ± 5.6*	91.2 ± 3.6*	91.4 ± 3.4*	89.6 ± 4.8*

Table 28
General average precision of the algorithms of classification in all databases using attribute selection (in %).

Algorithms						
Naïve Bayes	C4.5	SVM	1-NN	3-NN	5-NN	7-NN
83.0 ± 15.8	78.7 ± 13.3*	85.1 ± 14.2	78.2 ± 17.0*	79.47 ± 16.0	79.1 ± 15.9	79.2 ± 15.5

Table 29
Average precision on each database using the random projection method with fixed number of attributes (in %).

Databases	Algorithms						
	Naïve Bayes	C4.5	SVM	1-NN	3-NN	5-NN	7-NN
DLBCL	78. 0 ± 6.4	71.2 ± 4.4*	**82.3 ± 4.5**	69.8 ± 2.6*	72.1 ± 3.7*	73.6 ± 3.9*	74.6 ± 4.0*
DLBCL-Tumor	85.3 ± 3.4	81.3 ± 1.9*	**87.1 ± 6.9**	83.7 ± 3.6*	83.6 ± 1.7*	85.0 ± 2.5	85.4 ± 2.8
DLBCL-Outcome	43.3 ± 2.2	**48.8 ± 3.9**	**50.1 ± 2.3**	**53.7 ± 1.7**	**54.4 ± 0.7**	**52.2 ± 0.9**	**51.1 ± 1.7**
DLBCL-NIH	56.6 ± 1.2	55.0 ± 1.7*	**56.4 ± 0.4**	54.2 ± 1.0*	54.7 ± 1.2*	54.5 ± 1.3*	55.1 ± 1.5*
ALL/AML	86.2 ± 4.0	75.3 ± 2.0*	**88.3 ± 6.0**	83.5 ± 4.8*	84.3 ± 4.7*	83.3 ± 3.8*	82.1 ± 3.4*

Table 30
Average precision on each database using the random projection method with a percentage of the attributes (in %).

Databases	Algorithms						
	Naïve Bayes	C4.5	SVM	1-NN	3-NN	5-NN	7-NN
DLBCL	93.5 ± 1.8	75.3 ± 3.7*	93.0 ± 1.4	80.0 ± 2.3*	84.6 ± 3.1*	85.6 ± 2.7*	85.2 ± 2.9*
DLBCL-Tumor	88.5 ± 0.4	84.8 ± 1.6*	**96.4 ± 0.4**	87.9 ± 0.2*	86.8 ± 0.4*	**90.1 ± 0.9**	**91.0 ± 0.8**
DLBCL-Outcome	38.4 ± 1.3	**47.5 ± 1.3**	**52.0 ± 0.7**	**54.0 ± 2.0**	**56.8 ± 0.9**	**53.9 ± 1.2**	**50.9 ± 1.0**
DLBCL-NIH	60.0 ± 0.3	54.5 ± 1.4*	**62.4 ± 1.6**	56.9 ± 0.4*	60.0 ± 0.9	58.4 ± 0.8*	58.1 ± 0.5*
ALL/AML	92.7 ± 0.8	81.6 ± 1.1*	**97.4 ± 0.7**	91.5 ± 1.3*	92.1 ± 1.6*	90.8 ± 1.3*	89.9 ± 1.0*

Table 31
Average precision for on each database using the random projection method (in %).

Databases	Algorithms						
	Naïve Bayes	C4.5	SVM	1-NN	3-NN	5-NN	7-NN
DLBCL	85.8 ± 9.4	73.3 ± 4.4*	87.6 ± 6.4	74.9 ± 5.9*	78.6 ± 7.3*	79.6 ± 7.0*	79.9 ± 6.5*
DLBCL-Tumor	86.9 ± 2.9	83.1 ± 2.5*	**91.8 ± 6.7**	85.8 ± 3.3*	85.2 ± 2.1*	87.5 ± 3.2	**88.2 ± 3.5**
DLBCL-Outcome	40.9 ± 3.1	**48.2 ± 2.8**	**51.1 ± 1.9**	**53.8 ± 1.8**	**55.6 ± 1.5**	**53.0 ± 1.3**	**51.0 ± 1.3**
DLBCL-NIH	58.3 ± 2.0	54.7 ± 1.5*	**59.4 ± 3.4**	55.6 ± 1.6*	57.4 ± 3.0*	56.4 ± 2.3*	56.6 ± 1.9*
ALL/AML	89.5 ± 4.4	78.4 ± 3.7*	**92.9 ± 6.2**	87.5 ± 5.4*	88.2 ± 5.3	87.1 ± 4.8*	86.0 ± 4.7*

Table 32
General Average precision for each database using the random projection method (in %).

Algorithms						
Naïve Bayes	C4.5	SVM	1-NN	3-NN	5-NN	7-NN
72.2 ± 21.6	67.5 ± 15.3	76.5 ± 19.7	71.5 ± 16.1	72.9 ± 15.5	72.7 ± 16.8	72.3 ± 17.3

original database, one can observe that the results are statistically equivalent.

5. Conclusion

Some difficulty exists in defining which data mining technique is to be used on gene expression database obtained by the microarray technology. It happens because these databases request a differentiated treatment since they are composed by a great amount of attributes and relatively few samples. Under these circumstances, several machine learning algorithms have problems, as it is case of the artificial neuron networks. Then, dimensionality reduction methods can be seen as a solution in those types of data.

In this paper two methods of dimensionality reduction were applied: attribute selection and random projection method. Each reduction method was performed separately and the new set of attributes obtained. For attribute selection two approaches were used: the filter and wrapper.

Analyzing the results obtained by those methods it was observed a significant improvement in the results. When attribute selection algorithms were used, the success rate of the classifier was superior that obtained when applied on the database with all of the attributes, even in the worst cases. Besides that, there was a great reduction in the amount of attributes selected, mainly, when it was applied the evaluation measures combined with the sequential search.

It was observed that the wrapper approach, when applied together with the sequential search, produced better results, followed by the dependency evaluation measure with the filter approach. Although the difference of the result of the classifiers has been small regarding the success rate, the computational cost was very superior in relation to the use of the wrapper approach.

In general, the execution of the filter approach had a time of processing in the order of seconds to minutes. On the other hand, the use of the wrapper approach takes processing times in the order of hours to days, which can, in some cases, make unviable its application.

The SVM algorithm, in general, produced better results, however its processing time was quite high compared to other classification algorithms due to the size of the databases. Another algorithm that had good results in the classification was Naïve Bayes.

It was noticed a small increase in the classifiers' performance when using only the selected subset of attributes, regarding the use of all attributes. The SVM algorithm was the one that presented better results statistically and in none of the experiments it was considered worse than the base algorithm (Naïve Bayes).

Therefore, the general recommendation is to use attribute selection as a method of dimensionality reduction, since it produces good results when applied in gene expression database. The random projection method is an alternative method, because it reduces the computational cost and can provide better results than the original set of attributes.

Several extensions of this paper will now be explored. One of the proposals is to apply the random projection method with other dimensionality reduction methods to realize attribute selection. Another extension is the use of other classification algorithms, for instance, rule-based induction algorithm, since this type of algorithm can bring some surprising or unexpected knowledge.

References

Achlioptas, D. (2001). Database-friendly random projections. In *Proceedings of the ACM symposium on the principles of database systems*, (pp. 274–281).

Alizadeh, A. et al. (2000). Distinct types of diffuse large B-cell Lymphoma Identified by gene expression profiling. *Nature, 4051*, 503–511.

Bertoni, A., & Valentini, G. (2005). Random projections for assessing gene expression cluster stability, IJCNN '05. In *Proceedings IEEE international joint conference on neural networks*, 1, (pp. 149–154).

Borges, H. B., & Nievola, J. C. (2005). Attribute Selection Methods Comparison for Classification of Diffuse Large B-Cell Lymphoma. In *Proceedings of the fourth international conference on machine learning and applications–ICMLA05, Los Angeles*, 1, (pp. 201–206).

Borges, H. B. (2006). Redução de Dimensionalidade em Bases de Dados de Expressões Gênica. Dissertação (Mestrado em Informática Aplicada). 123f. PPGIA -Pontifícia Universidade Católica do Paraná - PUCPR.

Borges, H. B., & Nievola, J. C. (2008). Gene selection from microarray data. In Nadia Nedjah & Luiza M. Mourelle (Eds.). *Intelligent Text Categorization and Clustering* (vol. 164, pp. 1–23). Berlin–Heidelberg: Springer.

Borges, H. B., & Nievola, J. C. (2007). Gene-finding as an Attribute Selection Task. In *Proceedings of the sixth IEEE international conference on computer and information science–ICIS2007, 2007, Melbourne,* pp. 533–542.

Borges, H. B., & Nievola, J. C. (2009). Dimensionality Reduction in Gene Expression Database through the random projection method. In *Proceedings of the the eighth international conference on machine learning and applications–ICMLA 2009,* pp. 557–562.

Chuang Li-Yed, et al. (2008). *Computational Biology and Chemistry, 32*, 29–38.

Dash, M., & Liu, H. (1997). Feature Selection for Classification. *Intelligent Data Analysis: An International Journal, 1*(3), 131–156.

Fayyad, U., Haussler, D., & Stolorz, P. (1996). KDD for science data analysis: issues and examples. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, 1996 Portland, Oregon, Ago. AAAI Press*.

Goldberg, D. E. (1989). *Genetic algorithms in search, optimization, and machine learning*. Addison-Wesley Publishing Company.

Golub, T. et al. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science, 286*, 531–537.

Hall, M., 2000. Correlation-based feature selection for discrete and numeric class machine learning. In *Proceedings of the seventeenth international conference on machine learning,* pp. 359–366.

Kohavi, R., & John, G. H. (1998). The Wrapper Approach. In H. Liu, H. Motoda (Eds.) *Proceedings of the feature extraction, construction and selection: A data mining perspective,* (pp. 33–49).

Lin, J., & Gunopulos, D. (2003). Dimensionality Reduction by Random Projection and Latent Semantic Indexing. In *Proceedings of the data mining workshop, at the third SIAM international conference on data mining, San Francisco, CA*. Mai 3.

Liu, H., & Motoda, H. (1998). *Feature selection for knowledge discovery and data mining*. Kluwer Academic Publishers.

Liu, H., Motoda, H., & Yu, L. (2003). The Handbook of Data Mining, Lawrence Erlbaum Associates, Inc. Publishers. Editor: N.Y., (pp. 409–423).

Liu, H., & Yu, L. (2005). Toward integrating feature selection algorithms for classification and clustering. *IEEE Transactions on Knowledge and Data Engineering, 17*(4), 491–502.

Reynes Christelle, et al. (2008). *Computational Statistics and Data Analysis, 52*, 4380–4394.

Rosenwald, A. et al. (2002). The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *The New England Journal of Medicine, 346*(25), 1937–1947.

Shipp, M. et al. (2002). Diffuse large B-cell lymphoma outcome prediction by gene expression profiling and supervised machine learning. *Nature Medicine, 8*(1), 68–74.

Witten, I. H., Ian, H., & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques* (2nd ed.). San Francisco: Morgan Kaufmann.

Zhu Shenghuo, et al. (2010). Feature selection for gene expression using model-based entropy. *IEEE/ACM Transactions on Computational Biology and Bioinformatics, 7*(1).