American Scientific Research Journal for Engineering, Technology, and Sciences (ASRJETS)

ISSN (Print) 2313-4410, ISSN (Online) 2313-4402

© Global Society of Scientific Research and Researchers

http://asrjetsjournal.org/

Dimensionality Reduction Approach using Attributes Extraction and Attributes Selection in Gene Expression Databases

Helyane Bronoski Borges[a]*, Julio Cesar Nievola[b], Simone Nasser Matos[c], Rafael Felipe Tasaka de Melo[d], Raimundo Osvaldo Vieira[e]

[a,c,d,e]Federal University of Technology – Paraná (UTFPR), Address Doutor Washington Subtil Chueire, 330 - Jardim Carvalho, Ponta Grossa, 84017-220, Parana, Brazil

[b]Pontificia Universidade Catolica do Parana (PUCPR), Address Imaculada Conceição, 1155 - Prado Velho, Curitiba, 80215-901, Parana, Brazil

[a]Email: helyane@utfpr.edu.br, [b]Email: nievola@ppgia.pucpr.br, [c]Email: snasser@utfpr.edu.br, [d]Email: rafaelftmelo@outlook.com, [e]Email: raimundo.vieira@ifma.edu.br

Abstract

A high number of attributes form the gene expression databases. To deal with it, data dimensionality reduction is used to minimize the volume of data to be treated regarding the number of attributes and increase the generalization capability of learning methods by eliminating irrelevant and/or redundant data. This paper proposes an approach to means of dimensionality reduction, which joins attribute extraction and attributes selection. We used the Random Projection method, and the filter and wrapper approach for the attribute selection. The experiments are realized in five gene expression microarray databases. The results of the experiments showed that join of those approaches can provide promising results.

Keywords: Data Dimensionality Reduction; Attribute Selection; Attribute Extraction; Microarray.

1. Introduction

Dimensionality reduction consists of reducing the number of attributes to improve the performance of the classifiers [1]. In other words, reducing dimensionality is to find a significant representation in reduced dimensionality for high-dimensional data, maintaining a minimum number of parameters that preserve the properties observed in the data.

----

* Corresponding author.

This task is essential in several domains as it facilitates the classification, visualization, and understanding of the data [2]. The application dimensionality reduction methods provide some benefits such as reducing noise, redundancy between attributes, and finding a subset of relevant attributes [3]. Those methods are employed to improve the learning process's performance as the speed as in the classifier's performance [4]. 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 number of instances and the set of attributes. Those aspects are characteristics of gene expression databases obtained from the DNA microarray. Data are formed by a very high number of attributes (genes) and a small number of instances (samples). Several approaches to data dimensionality reduction have been proposed and evaluated [5]. To reduce the dimensionality, there are two basic approaches: attribute extraction and attribute selection [5]. Attribute extraction algorithms create new attributes starting from transformations or combinations of the original group of attributes. As their name suggests, selection algorithms select according to a specific criterion the best subset of the original group of attributes. Moreover, the application of the attributes selection and attribute extraction in gene expression database can still increase the generated results' comprehensibility, identifying the influence of each selected attribute. This paper proposes a dimensionality reduction approach, which joins random projection with attributes selection in five gene expression databases with similar characteristics. In the attribute selection used: the filter and the wrapper approach. To the attribute extraction applied the Random Projection. The proposed approach is compared with others in the literature.

2. Dimensionality Reduction

The literature presents some dimensionality reduction methods of the two approaches: Attribute Selection and Attribute Extraction [5]. The main difference between the tree structure and the DAG structure is that in the tree structure each node (each class), except the root node, has only one ancestor (parent), while in the DAG structure each node (class) may have one or more ancestors' nodes.

2.1. Attribute Selection

Through the attribute selection, a subset of M attributes out of the N original attributes is chosen, such as $M < N$, in a way that the number of the attributes is reduced according to a pre-established condition [6]. Attribute selection tries to guarantee that the data into the mining stage has a good quality. The subset generation is a search procedure that creates subsets of candidate attributes 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. But if the new subset is worse, 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 through different test samples obtained from a real or synthetic dataset [6, 7]. This process's nature depends on two basic aspects. First, one should decide the search starting point (or points), determining the search direction. The search can start with an empty set and steadily add attributes to the existing set (forward search), or it can start with a complete set of attributes and remove one attribute at each iteration (backward search). Another possibility is to start with a predefined set of attributes and add/remove attributes in each step [8]. When the search begins with a randomly selected subset of attributes, it can avoid the local minimum trap. The second decision to be made regards the search strategy. For a dataset with *N* attributes, there are *2N* possible subsets. This search room grows exponentially, making it prohibitive even for a moderate value of *N*. According to this line of reasoning, search algorithms can be divided into three main groups: exponential, sequential, and random algorithms [7]. Each subset created needs to be evaluated using the evaluation criterion. A subset's quality can be computed according to a specific criterion (for instance, a selected optimum subset defined by using one criterion could not be optimal according to another criterion). In a general way, the evaluation criteria can be categorized into two groups: filter and wrapper. The filter approach belongs to the independent criterion [9, 10]. It tries to evaluate an attribute (or attributes) subset exploring the training data's intrinsic features without any commitment to the mining algorithm. The most used independent criteria are: distance measures, information measures, dependency measures, and consistency measures [9]. The dependent criterion characterizes the wrapper approach [10]. It requires a predefined mining algorithm within the attribute selection, and it uses its performance when applied in the selected subset to evaluate the quality of its attributes. The stopping criterion establishes when the attribute selection process should be finished. It can be done when the search is over or when the goal is reached, where the goal can be a specific situation (maximum number of characteristics or maximum number of iterations), or when a good enough 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. Attribute Extraction

The extraction attribute is an approach where new attributes are created from the original database through linear and nonlinear combinations [5]. The objective is to make them more expressive and to represent better the variability of the data. Attribute extraction can be defined as a set *N* formed by n attributes which, after undergoing an attribute extraction process, generates a new set *N*, with *m* attributes such where $m < n$. Thus, a new feature extracted is obtained by the $f_m(N)$, where *f* is a mapping function that processes linear or nonlinear transformations on the original attribute set [1]. Linear transformations do not modify the data's spatial structure, preserving the relationships between the observations [11]. The Principal Component Analysis (PCA) is an example of the Linear Transformations Algorithm [12], whose purpose is to reduce data on which the variables are highly dependent, maintaining maximum variability. For this, it generates new attributes as linear functions of the original attributes. In nonlinear problems, attribute extraction often involves the application of nonlinear transformations. These transformation methods are efficient when approximating functions and robust when handling real nonlinear problems [13]. Nonlinear transformations modify (increase or decrease) the linear relationships between variables by changing their correlation. For example, according to Bingham and Mannila in [14], the Random Projection Method is a nonlinear method in which the original high-dimensional data is projected into a smaller sub-space using a random matrix.

2.3. Related Works

Many works related to dimensionality reduction can be found in the literature using attribute selection algorithms [3, 15, 16, 17, 18] and attribute extraction algorithms [19, 20] applied in microarray data. However, few works apply both methods extraction and selection attribute together. Cui and his colleagues in [21] proposed a method named Dispersed Maximum Margin Discrimination Analysis (SMMDA) based on the Maximum Margin Criterion (MMC) - becoming an extension of the LDA. In this method, a sparse representation of the data is used to replace the K-nearest neighbor technique. Unlike MMC, the method does not need the weighting function to deemphasize the samples far from the classification margin. To perform the gene selection, the SMMDA is applied. Then, for each gene, a calculation is made, generating a score. The genes are sorted based on their scores in descending order, and the first genes are selected. Five different databases were used for tests from the research (colon, leukemia, glioma, prostate, and breast), and the proposed method was compared to other extraction algorithms (PCA, LDA, and PLS), leading to results in which the method is considered to be efficient when extracting discriminatory gene characteristics. You and his colleagues in [22] proposed a new local dimension reduction algorithm named TotalPLS, which operates in a unified partial least squares (PLS) framework and implements an information fusion of PLS-based feature selection and feature extraction. Eight cancer microarray datasets were used for the experiments. The researchers compared the results in terms of recognition accuracy, relevance, and redundancy. The Multi-Filtration Feature Selection (MFFS) was applied on 22 different medical databases in four stages: first, the PCA (Principal Component Analysis) method was proposed by Sasikala and his colleagues in [23] to extract the most relevant attributes. Then, it was applied the CFS (Correlation-based Feature Selection) to select the features. In the third stage, the attribute was evaluated using the Symmetrical Uncertainty (SU). In the fourth stage, classifiers algorithms were applied. The results were rebuilding those ranks to ensure that any overvaluation had been withdrawn and, lastly, ranking. The method has become quite effective among other selection methods and has even presented the best precision. Badaoui and his colleagues in [24] proposed a processing approach that is structured on feature extraction and selection. The attribute extraction uses Linear Discriminant Analysis Method (LDA), while the selection consists of eliminating redundant or irrelevant variables using some adapted techniques of discriminant analysis. The approach is tested on three microarray databases. The dimensionality reduction technique based on the Fuzzy-Rough Theory was proposed by [25]. Depending on the criterion to be adopted by the algorithm, it can extract and select characteristics simultaneously. First, a calculation is done to know how important a characteristic is - erasing redundant or insignificant data through mathematical calculation. Subsequently, using the extraction or selection of features, a new low base is generated (if the new generation base is not enough, new characteristics are generated, or some are removed). For the tests, many bases were used (two bases of breast cancer, two of leukemia, sat image, colon cancer, lung cancer, isolet, multiple features, and segmentation) and applied on classification algorithms (SVM, KNN, and Decision Tree C4.5). The results showed that the proposed method is efficient compared to algorithms with the same proposal.

3. Methodology of Experiments

The proposed approach consists of the joint of the attributes extraction (Random Projection) and attributes selection (Filter and Wrapper) to reduce dimensionality, as illustrated in Figure 1. In this work, five microarray databases (see Table 1) were submitted to four classifiers: Naïve Bayes, C4.5, SVM, and k-NN (for k=1, k=3, k=5, and k=7) [4] as a way to compare the classification rate of the classifier using all attributes.

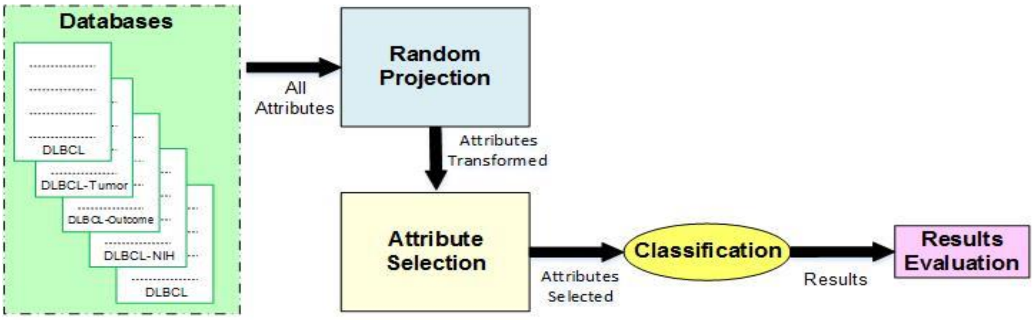


Figure 1: General Approach to Dimensionality Reduction.

The databases are the input for the dimensionality reduction process. After all attributes from databases are applied on the Random Projection, it generates the transformed ones. These attributes are applied to the attribute selection algorithms. The new subsets are generated and used on the classification algorithms for evaluation results.

Table 1: Characteristics of Databases

Databases	Samples	Attributes
DLBCL	47	4026
DLBCL-Tumor	77	7129
DLBCL-Outcome	58	7129
DLBCL-NIH	71	7129
AML-ALL	240	7399
DLBCL	47	4026

These attributes are applied to attribute selection algorithms. The new attribute subsets are generated and used on the classification algorithms for evaluation of the results. Figure 2 shows details of Stage 1 to the proposed application approach. The new attributes group's dimension was defined in a random way for the five databases using the Random Projection. The database represents all the databases, shown in Figure 4, that had Random Projection Approach applied. This approach uses fixed number attributes (10, 15, 30, 45, 71) and percentage attributes (3%, 10%, 20%, 25%, 50%) where each value (1.1, 1.2,...,5.10; 6.1,6.2,...,10.10) represents a subset from the selected database. The results of the approach are 100 attributes subsets transformed. The development of the subsets was based on Borges and Nievola (see in [16]). Achlioptas described the distribution used for the calculation of the random matrix in [26]. Stage 2 from the proposed approach is illustrated in Figure 3. The 100 attribute subsets generated by Stage 1 were submitted to the attribute selection algorithms. Thus, for each subset were generated 18 new subsets. For example, considering the subset 1.1 generates 1.1 S+D...,1.1 R+W(7-NN). These processes are repeated for the 100 subsets, totaling 1800 subsets submitted to classification algorithms to evaluate the results.

4. Results and Discussion

The research objective was to compare the results from different combinations of search methods and evaluation criteria of the generated subsets using some classifiers in five microarray databases.   The results were obtained by running algorithms on the original database and the attributes subsets generated by the chosen dimensionality reduction method, which means combining attribute selection and random projection. The five chosen databases were submitted to the four classifiers. On the tables containing the results, the note "*" indicates that a result statistically is significantly worse than the result of the standard algorithm (Naive Bayes), and the result in bold indicates that it is significantly better compared to the standard algorithm.

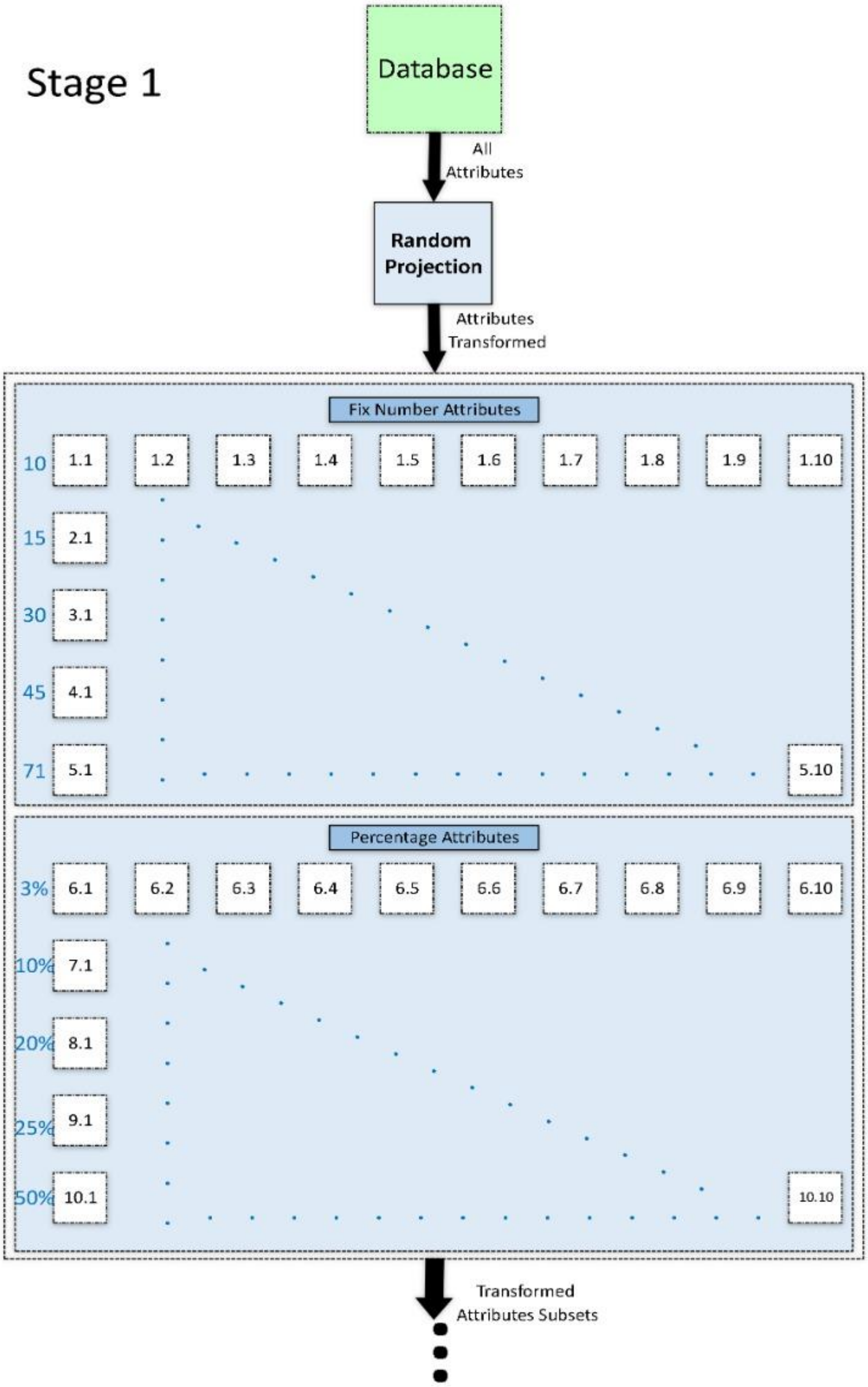
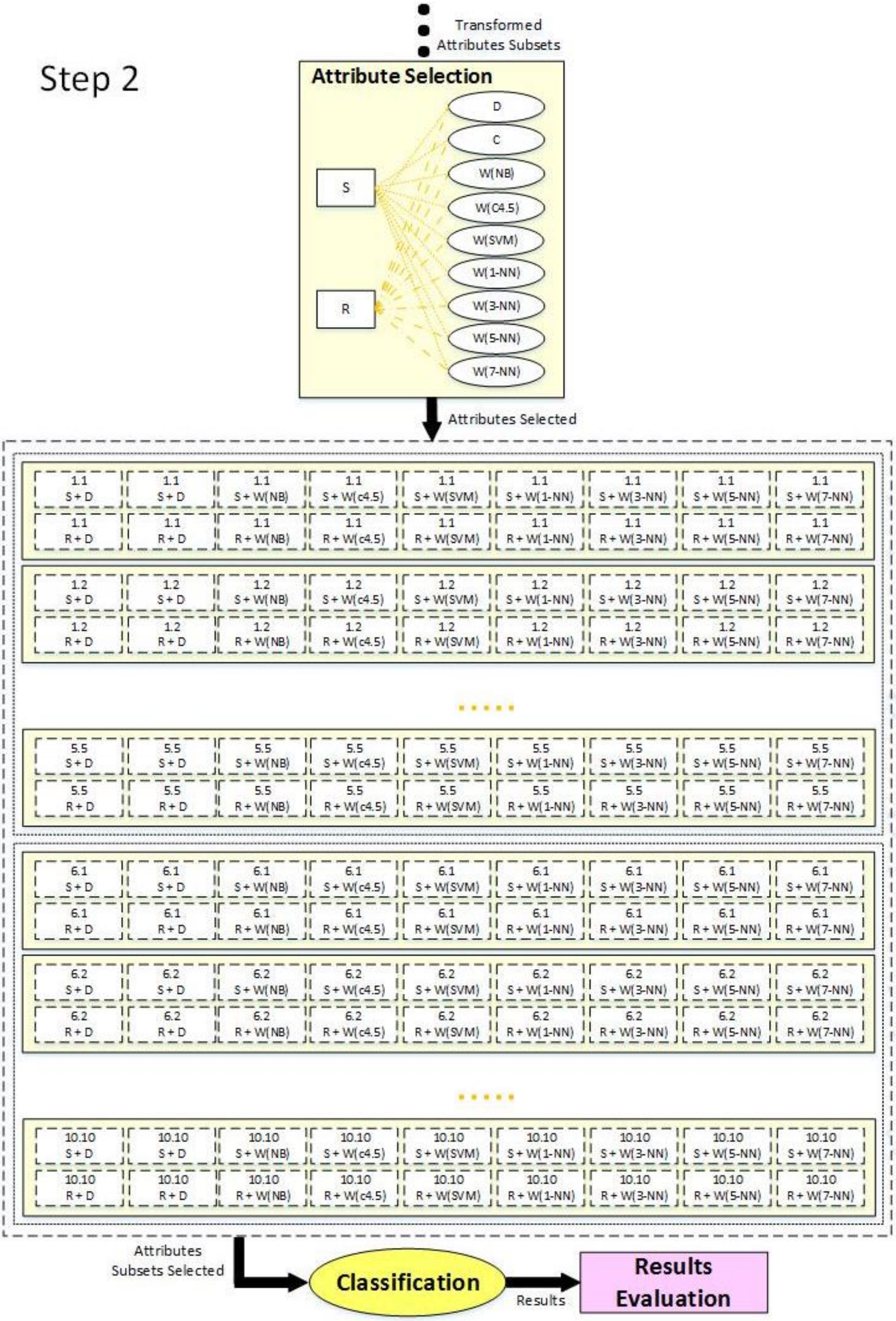


Figure 2: Stage 1 - Detail Approach to Dimensionality Reduction.



1.1 S + D	1.1 S + D	1.1 S + W(NB)	1.1 S + W(c4.5)	1.1 S + W(SVM)	1.1 S + W(1-NN)	1.1 S + W(3-NN)	1.1 S + W(5-NN)	1.1 S + W(7-NN)
1.1 R + D	1.1 R + D	1.1 R + W(NB)	1.1 R + W(c4.5)	1.1 R + W(SVM)	1.1 R + W(1-NN)	1.1 R + W(3-NN)	1.1 R + W(5-NN)	1.1 R + W(7-NN)
1.2 S + D	1.2 S + D	1.2 S + W(NB)	1.2 S + W(c4.5)	1.2 S + W(SVM)	1.2 S + W(1-NN)	1.2 S + W(3-NN)	1.2 S + W(5-NN)	1.2 S + W(7-NN)
1.2 R + D	1.2 R + D	1.2 R + W(NB)	1.2 R + W(c4.5)	1.2 R + W(SVM)	1.2 R + W(1-NN)	1.2 R + W(3-NN)	1.2 R + W(5-NN)	1.2 R + W(7-NN)

.....

5.5 S + D	5.5 S + D	5.5 S + W(NB)	5.5 S + W(c4.5)	5.5 S + W(SVM)	5.5 S + W(1-NN)	5.5 S + W(3-NN)	5.5 S + W(5-NN)	5.5 S + W(7-NN)
5.5 R + D	5.5 R + D	5.5 R + W(NB)	5.5 R + W(c4.5)	5.5 R + W(SVM)	5.5 R + W(1-NN)	5.5 R + W(3-NN)	5.5 R + W(5-NN)	5.5 R + W(7-NN)
6.1 S + D	6.1 S + D	6.1 S + W(NB)	6.1 S + W(c4.5)	6.1 S + W(SVM)	6.1 S + W(1-NN)	6.1 S + W(3-NN)	6.1 S + W(5-NN)	6.1 S + W(7-NN)
6.1 R + D	6.1 R + D	6.1 R + W(NB)	6.1 R + W(c4.5)	6.1 R + W(SVM)	6.1 R + W(1-NN)	6.1 R + W(3-NN)	6.1 R + W(5-NN)	6.1 R + W(7-NN)
6.2 S + D	6.2 S + D	6.2 S + W(NB)	6.2 S + W(c4.5)	6.2 S + W(SVM)	6.2 S + W(1-NN)	6.2 S + W(3-NN)	6.2 S + W(5-NN)	6.2 S + W(7-NN)
6.2 R + D	6.2 R + D	6.2 R + W(NB)	6.2 R + W(c4.5)	6.2 R + W(SVM)	6.2 R + W(1-NN)	6.2 R + W(3-NN)	6.2 R + W(5-NN)	6.2 R + W(7-NN)

.....

10.10 S + D	10.10 S + D	10.10 S + W(NB)	10.10 S + W(c4.5)	10.10 S + W(SVM)	10.10 S + W(1-NN)	10.10 S + W(3-NN)	10.10 S + W(5-NN)	10.10 S + W(7-NN)
10.10 R + D	10.10 R + D	10.10 R + W(NB)	10.10 R + W(c4.5)	10.10 R + W(SVM)	10.10 R + W(1-NN)	10.10 R + W(3-NN)	10.10 R + W(5-NN)	10.10 R + W(7-NN)

Figure 3: Example of the training process.

The results are presented in three subsections: results of all the attributes, results of the proposed approach, and a general comparison of the results. The individual's fitness is evaluated using the measure approach Distance-based Depth-Dependent Measures. When evaluating the result of a hierarchical prediction, three situations may occur: correct prediction, partially correct prediction, and incorrect prediction. Each of these situations will be exemplified.

4.1. Results with all attributes

Table 2 shows the results when applied the classifiers on the data using all attributes in five databases. It is observed that the 1-NN, 3-NN, and 7-NN algorithms in the DLBCL database attribute subsets had statistically worse results. Although the 3-NN algorithm has a better result than the 5-NN algorithm and has a hit rate equal to the C4.5 algorithm, it is considered statistically worse. In the DLBCL-Tumor database's attributes subsets, the SVM algorithm showed better results than the base algorithm. Although the C4.5 algorithm has a low hit rate, it is considered equivalent to the Naïve Bayes algorithm.

Table 2: Results of the databases' classification with all the attributes (precision, in %).

Databases	Naive 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
AML-ALL	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*

DLBCL-Outcome and DLBCL-NIH databases' hit rates were low compared to other databases, but these results are considered statistically equivalent. Analyzing the results from the ALL/AML database, the 7-NN algorithm has a lower hit rate. The C4.5, 3-NN, and 5-NN algorithms also have worse results.

4.2. Results obtained with the proposed approach

This section presents the results after applying the proposed new approach to classify attribute subsets using Filter and Wrapper.

Filter Approach

The same criterion was used to execute attribute selection and classification. It is not possible to present the

1800 results obtained in each database. First, an arithmetic calculation of all ten executions of each attribute subset generated by the random projection method was calculated. We obtained the results from each attribute selection method of each attribute subset. After that, the manner was again determined from all of the previously obtained results, and these are the results presented. Table 3 presents the joint use of the random projection method and the selection of attributes in the DLBCL database. It is observed that in subset 1.1, the algorithms C4.5 and k-NN, for the values of k = 1, k = 3, and k = 5, were considered statistically worse when compared with the Naïve Bayes base algorithm. In subset 1.2, only the 1-NN algorithm is statistically worse than Naïve Bayes. The others are considered equivalent. Still, in subsets 2.1 and 2.2, only the SVM algorithm is considered equivalent to the base algorithm; the rest is statistically worse.

Table 3: Average precision on each database using the joint use of Random Projection Method and Attributes Selection using the Filter Approach in the DLBCL Subsets (in %).

Datasets	Naïve Bayes	C4.5	SVM	1-NN	3-NN	5-NN	7-NN
S + D	87,70 ± 12,17*	77,30 ± 4,99	87,79 ± 11,45	83,62 ± 12,38*	85,62 ± 11,71*	85,95 ± 12,13*	86,43 ± 12,06
S + C	83,09 ± 8,23	80,96 ± 7,55	82,89 ± 8,24	80,31 ± 10,26*	82,48 ± 9,40	82,08 ± 9,42	82,66 ± 9,51
R + D	83,00 ± 12,20	73,17 ± 7,29*	83,62 ± 11,91	75,03 ± 10,03*	77,83 ± 10,17*	79,14 ± 10,72*	79,95 ± 10,92*
R + C	82,26 ± 7,03	72,57 ± 3,15*	82,76 ± 7,18	72,80 ± 5,27*	76,33 ± 6,29*	77,25 ± 6,09*	77,18 ± 6,00*

Comparing the results from the Dimensionality Reduction Method with the results of the algorithms in the original database (Table 2), it is observed that in only two cases, in the Naïve Bayes algorithm and the SVM, the result without applying the Method of Reduction of Dimensionality was better. Table 4 presents the joint use of the random projection method and the selection of attributes in the DLBCL-Tumor database. In these experiments, the C4.5 algorithm had its results considered statistically worse than the Naïve Bayes base algorithm in all cases. In addition to the C4.5 algorithms, the 1-NN algorithm in the subsets S+D and R+C and the SVM algorithm in the subset S+C had their results statistically worse in their subsets. It is worth mentioning the 7-NN algorithm that had statistically better results than the base algorithm in the subsets of S+C, R+D, and R+D attributes.

Table 4: Average precision on each database using the joint use of Random Projection Method and Attributes Selection using the Filter Approach in the DLBCL-Outcome Subsets (in %).

Datasets	Naïve Bayes	C4.5	SVM	1-NN	3-NN	5-NN	7-NN
S + D	63,29 ± 8,64	60,74 ± 5,79*	60,34 ± 6,18*	59,98 ± 5,69*	61,28 ± 6,18*	61,28 ± 6,33*	61,13 ± 6,77*
S + C	62,03 ± 6,81	61,90 ± 7,52	59,62 ± 4,82*	59,05 ± 6,03*	61,43 ±6,30*	61,00 ± 6,12*	61,27 ± 6,77
R + D	23,29 ± 18,28	**29,43 ± 24,95**	**27,71 ± 23,11**	**29,64 ± 25,52**	**30,00 ± 25,42**	**29,31 ± 25,12**	**28,09 ± 24,14**
R + C	45,07 ± 6,25	**53,29 ± 1,39**	**51,83 ± 3,35**	**52,60 ± 2,89**	**52,71 ± 3,42**	**53,31 ± 1,46**	**51,90 ± 1,77**

If we compare these results with those obtained when using the database with all of the attributes, it is noticed that when the SVM and 3-NN algorithm were used, the results were lower, but the rest of the results were at least equal. Analyzing the results obtained from the joint use of the random projection method and the attributes selection in the subsets of attributes of the DLBCL-Outcome database (Table 4), it is observed that all the results of the algorithms of the subset S+D are considered statistically worse. This also happens for the SVM, 1-NN, 3-NN, and 5-NN algorithms in the S+C subset. Yet, the results of the subsets R+D and R+C algorithms are statistically better than the Naïve Bayes base algorithm. Table 5 presents the results obtained from the random projection method's joint use and the attributes selection in the attribute subsets of the DLBLC-NIH database. By defining the Naïve Bayes algorithm as the base algorithm to be compared statistically with the other algorithms for the subset of S+D attributes, all other algorithms are considered worse. For the subset of S+D attributes, only the result of the algorithm C4.5 is considered equivalent to the Naïve Bayes one. For the subsets of R+D and R+C attributes, only the result of the SVM is considered equivalent to the result of the Naïve Bayes base algorithm; the rest of the results are statistically worse.

Table 5: Average precision on each database using the joint use of Random Projection Method and Attributes Selection using the Filter Approach in the DLBCL-NIH (in %).

Datasets	Naïve Bayes	C4.5	SVM	1-NN	3-NN	5-NN	7-NN
S + D	62,95 ± 5,48	59,95 ± 1,77*	61,10 ± 4,11*	55,93 ± 2,52*	57,29 ± 3,65*	58,22 ± 4,45*	58,48 ± 4,83*
S + C	63,28 ± 5,65	60,64 ± 2,40	61,52 ± 4,68*	56,27 ± 3,30*	57,47 ± 4,50*	57,89 ± 5,40*	58,12 ± 5,73*
R + D	41,58 ± 23,07	40,04 ± 21,44*	42,05 ± 23,75	38,45 ± 21,40*	39,21 ± 22,03*	39,98 ± 21,91*	39,92 ± 22,05*
R + C	59,08 ± 2,01	56,40 ± 1,66*	59,21 ± 1,87	54,23 ± 2,98*	55,55 ± 3,87*	56,05 ± 2,92*	56,32 ± 2,32*

Table 6 shows the combined use of dimensionality reduction methods in the ALL/AML database attribute subsets. In the subsets of S+D and R+D attributes, the algorithms C4.5 and k-NN are statistically worse than the

Naïve Bayes algorithm.

Table 6: Average precision on each database using the joint use of Random Projection Method and Attributes Selection using the Filter Approach in the ALL/AML (in %).

Datasets	Naïve Bayes	C4.5	SVM	1-NN	3-NN	5-NN	7-NN
S + D	93,26 ± 6,69	81,86 ± 4,34*	92,37 ± 7,77	90,35 ± 9,82*	90,56 ± 8,96*	90,72 ± 8,48*	91,01 ± 8,1*
S + C	88,82 ± 4,98	85,24 ± 5,96*	87,45 ±5,74*	85,40 ± 7,80*	86,43 ± 6,93*	86,75 ± 6,27*	86,69 ± 6,10*
R + D	89,23 ± 6,35	77,95 ± 4,56*	89,68 ± 8,64	86,42 ± 8,95*	87,05 ± 8,19	86,84 ± 7,49*	86,70 ± 6,97*
R + C	87,01 ± 3,78	77,74 ± 1,24*	**88,96 ± 6,19**	85,23 ± 5,18*	85,90 ± 5,85	85,03 ± 4,96*	84,06 ± 4,17*

In the subset S+C besides the algorithms C4.5 and k-NN, the SVM algorithm's result is also statistically worse. In the subset of R+C attributes, the SVM algorithm has a better result than the base algorithm and the algorithms C4.5 and k-NN, for k = 1, k = 5, and k = 7.

Wrapper Approach

The search method identified with S refers to the Sequential Search, and the search with R refers to the Random Search. Table 7 refers to the results obtained from the Random Projection Method's joint use and the attribute selection in the DLBCL base attribute subsets. Analyzing the search method S, it is observed that the algorithms C4.5, 1-NN, and 3-NN are statistically worse than the Naïve Bayes algorithm. However, for the search method R, the algorithms C4.5 and the k-NN (for the values of k = 1, k = 3, k = 5, and k = 7) are statistically better.

Table 7: Average precision on each database using the joint use of Random Projection Method and Attributes Selection using the Wrapper Approach in the DLBCL Subsets (in %).

Datasets	Naïve Bayes	C4.5	SVM	1-NN	3-NN	5-NN	7-NN
S	91,16 ± 7,67	85,72 ± 6,82*	91,32 ± 6,60	89,72 ± 9,03*	89,63 ± 9,02*	89,62 ± 9,64	90,83 ± 7,61
R	92,19 ± 6,93	82,88 ± 4,34*	**94,38 ± 6,19**	88,61 ± 5,16*	89,42 ± 4,16*	89,55 ± 4,06*	89,70 ± 4,39*

Table 8 refers to the results obtained from the random projection method's joint use and the attribute selection in the DLBCL-Tumor base attribute subsets. It is observed that the results of the C4.5 algorithm in these two search methods were statistically worse compared to the base algorithm and the result of the 5-NN algorithm was statistically better in the Sequential Search Method. The result of the C4.5 algorithm was statistically worse than the Naïve Bayes algorithm in two search experiments. The 5-NN algorithm had its result statistically better in the search method S.

Table 8: Average precision on each database using the joint use of Random Projection Method and Attributes Selection using the Wrapper Approach in the DLBCL-Tumor Subsets (in %).

Datasets	Naïve Bayes	C4.5	SVM	1-NN	3-NN	5-NN	7-NN
S	92,78 ± 6,06	88,86 ± 6,91*	88,40 ± 10,47	94,58 ± 6,66	94,68 ± 4,72	95,09 ± 4,74	94,60 ± 5,43
R	92,09 ± 2,84	88,60 ± 5,78*	91,77 ± 11,23	92,34 ± 7,74	91,13 ± 7,58	91,91 ± 7,58	92,99 ± 5,01

Table 9 shows the results obtained from the DLBCL-Outcome database in which the k-NN algorithm had better results statistically compared to Naïve Bayes.

Table 9: Average precision on each database using the joint use of Random Projection Method and Attributes Selection using the Wrapper Approach in the DLBCL-Outcome Subsets (in %).

Datasets	Naïve Bayes	C4.5	SVM	1-NN	3-NN	5-NN	7-NN
S	71,48 ± 7,13	69,93 ± 12,46	70,28 ± 10,61	**75,17 ± 7,29**	**75,22 ±6,81**	**74,40** ± **7,29**	72,66 ± 8,20
R	55,45 ± 8,54	58,12 ± 1,81	62,10 ± 4,03	**69,10 ± 3,24**	**68,38 ±2,59**	**65,71** ± **3,52**	**63,81 ± 3,45**

Table 10 shows the results obtained from the DLBCL-NIH database. The k-NN algorithm had the worst results statistically evaluated when using the search method S. Although, in the search method R, the algorithm C4.5 had a worse result. Table 11 shows the results obtained from the ALL/AML database in which the C4.5 algorithm had worse results than the Naïve Bayes algorithm in the two search methods. The 5-NN and 7-NN algorithms also had statistically worse results in the Random Search Method.

Table 10: Average precision on each database using the joint use of Random Projection Method and Attributes Selection using the Wrapper Approach in the DLBCL-NIH Subsets (in %).

Datasets	Naïve Bayes	C4.5	SVM	1-NN	3-NN	5-NN	7-NN
S	65,74 ± 3,72	66,48 ± 7,10	63,41 ± 6,91	63,90 ± 2,60*	64,44 ± 3,18*	64,69 ± 2,92*	65,05 ± 3,00*
R	63,68 ± 1,42	58,32 ± 1,06*	64,58 ± 5,01	63,96 ± 0,89	64,17 ± 1,29	63,92 ± 0,86	63,48 ± 1,29

Table 11: Average precision on each database using the joint use of Random Projection Method and Attributes Selection using the Wrapper Approach in the ALL/AML Subsets (in %).

Datasets	Naïve Bayes	C4.5	SVM	1-NN	3-NN	5-NN	7-NN
S	93,97 ± 5,13*	88,75 ± 6,21	92,96 ± 7,17	93,55 ± 6,20	93,58 ± 5,37	91,37 ± 6,89	91,45 ± 6,85
R	94,33 ± 4,71*	86,48 ± 3,54	94,72 ± 5,89	94,47 ± 4,20	93,95 ± 3,85	91,91 ± 4,56*	91,06 ± 4,47*

Table 12 presented the general average of the classification results process where the Wrapper Approach obtained the best results compared to Filter in all databases.

Table 12: Comparison between the Attributes Selection Methods when applied together with the Random Projection method in the five databases (in %).

Subsets	DLBCL	DLBCL-Tumor	DLBCL-Outcome	DLBCL-NIH	ALL/AML
S + D	84,92	90,30	61,15	59,13	90,02
S + C	82,07	88,10	60,90	59,31	86,68
R + D	78,82	86,73	28,21	40,18	86,27
R + C	77,31	86,16	51,53	56,69	84,85
S + W (NB)	91,16	92,78	71,48	65,74	93,97
S + W (C4.5)	85,72	88,86	69,93	66,48	88,75
S + W (SVM)	91,32	88,40	70,28	63,41	92,96
S + W (1-NN)	89,72	94,58	75,17	63,90	93,55
S + W (3-NN)	89,63	94,68	75,22	64,44	93,58
S + W (5-NN)	89,62	95,09	74,40	64,69	91,37
S + W (7-NN)	90,83	94,60	72,66	65,05	91,45
R + W (NB)	92,19	92,09	55,45	63,68	94,33
R + W (C4.5)	82,88	88,60	58,12	58,32	86,48
R + W (SVM)	94,38	91,77	62,10	64,58	94,72
R + W (1-NN)	88,61	92,34	69,10	63,96	94,47
R + W (3-NN)	89,42	91,14	68,38	64,17	93,95
R + W (5-NN)	89,55	91,91	65,71	63,92	91,91
R + W (7-NN)	89,70	92,99	63,81	63,48	91,06

4.3. General Comparison

Table 13 shows the average of the executions of the approach proposed, using the filter approach of each database. It is observed that in all databases, the SVM algorithm had a statistically equivalent result to the base algorithm. It is also noticed that the algorithms C4.5 and k-NN had statistically worse results on the average of almost all databases.

Table 13: Average precision for each database using the proposed approach - Filter Approach (in %).

Databases	Naïve Bayes	C4.5	SVM	1-NN	3-NN	5-NN	7-NN
DLBCL	84,0 ± 2,5	76,0 ± 3,9*	84,3± 2,4	77,9 ± 4,9*	80,57 ± 4,3*	81,1 ± 3,8*	81,6 ± 4,0*
DLBCL-Tumor	88,6 ± 2,43	84,3 ± 2,4*	88,1 ± 2,8	87,0 ± 2,2*	88,4 ± 2,4	88,7 ± 2,1	**89,6 ± 1,8**
DLBCL-Outcome	48,4 ± 18,7	51,4 ± 15,1	49,9 ± 15,3	50,3 ± 14,2	51,4 ± 14,8	51,22± 15,1	50,6 ± 15,6
DLBCL-NIH	56,7 ± 10,3	54,7 ± 9,7*	56,0 ± 9,3	51,2 ± 8,6*	52,4 ± 8,8*	53,0 ± 8,8*	53,2 ± 8,9*
ALL/AML	89,6 ± 2,6	80,7 ± 3,6*	89,6 ± 2,1	86,9 ± 2,4*	87,5 ± 2,1*	87,4 ± 2,4*	87,1 ± 2,9*

Table 14 shows the average of the approach's executions proposed with the attribute selection using the Wrapper Approach for each database. In the DLBCL database, the SVM algorithm had statistically better results, and the other algorithms had worse ones than the base algorithm. DLBCL-Tumor database and the 1-NN algorithm had statistically better results, and the C4.5, SVM, 5-NN, and 7-NN algorithm results were statistically worse than the Naïve Bayes algorithm. DLBCL-Outcome database, the k-NN algorithm presented better results, and the C4.5 algorithm had the worst result. In the DLBCL-NIH database, all of the algorithms results were statistically worse than the Naïve Bayes algorithm. Also, in the ALL/AML database, the C4.5 algorithm and the k-NN presented the worst results statistically compared with the base algorithm. Table 15 shows the general average of the application of the Random Projection Method of each database. Table 16 shows the classification algorithms' overall position using the Random Projection Method's joint use and the Attribute Selection. Analyzing the results obtained and carrying out the statistical analysis, all results are statistically equivalent compared with the Naïve Bayes algorithm.

Table 14: Average precision for each database using the approach proposed - Wrapper Approach (in %).

Databases	Naïve Bayes	C4.5	SVM	1-NN	3-NN	5-NN	7-NN
DLBCL	91,7 ± 0,8	84,3 ± 20,1*	**92,9 ± 2,2**	89,2 ± 0,8*	89,5 ± 0,2*	89,9 ± 0,1*	90,3 ± 0,8*
DLBCL-Tumor	92,4 ± 0,5	88,7 ± 0,2*	90,1 ± 2,4*	**93,5 ± 1,6**	92,9 ± 2,5	93,5 ± 2,3*	93,8 ± 1,1*
DLBCL-Outcome	63,5 ± 11,3	64,0 ± 8,4*	66,2 ± 5,8	**72,1 ± 4,3**	**71,8 ± 4,8**	**70,1 ± 6,1**	**68,2 ± 6,3**
DLBCL-NIH	64,7 ± 1,5	62,4 ± 5,8*	64,0 ± 0,8*	63,9 ± 0,1*	64,3 ± 0,2*	64,3 ± 0,6*	64,3 ± 1,1*
ALL/AML	94,2 ± 0,3	87,6 ± 1,6*	93,9 ± 1,3*	94,0 ± 0,6	93,8 ± 0,3*	91,6 ± 0,4*	91,3 ± 0,3*

Table 15: General Average classification algorithm in all databases using the approach proposed (in %).

Databases	Naïve Bayes	C4.5	SVM	1-NN	3-NN	5-NN	7-NN
DLBCL	86,6 ± 4,4	78,8 ± 5,3*	87,1 ± 4,9	81,7 ± 7,0*	83,6 ± 5,7*	83,9 ± 5,7*	84,5 ± 5,5*
DLBCL-Tumor	89,9 ± 2,7	85,8 ± 2,9*	88,8 ± 2,6*	89,1 ± 3,8	89,9 ± 3,2	90,3 ± 3,1	**91,0 ± 2,6**
DLBCL-Outcome	53,4 ± 17,2	55,6 ± 13,9	55,3 ± 14,8	57,6 ± 15,9	58,2 ± 15,7	57,5 ± 15,4	56,5 ± 15,4
DLBCL-NIH	59,4 ± 9,0	**57,0 ± 9,0**	58,6 ± 8,3	55,5 ± 9,3*	56,4 ± 9,2*	56,8 ± 9,0*	56,9 ± 9,0*
ALL/AML	91,1 ± 3,1	83,0 ± 4,6*	91,1 ± 2,8	89,2 ± 4,2*	89,6 ± 3,6*	88,8 ± 2,9*	88,5 ± 3,1*

Table 16: General Average classification algorithm using the approach proposed (in %).

Naïve Bayes	C4.5	SVM	1-NN	3-NN	5-NN	7-NN
76,1 ± 18,1	72 ± 14,6	76,2 ± 17,6	74,6 ± 16,8	75,5 ± 16,9	75,5 ± 16,9	75,5 ± 17,3

We also applied only an attribute selection and a random projection in the same database. Here we compared these results with the proposed approach. Table 17 presents the two general averages for all attributes (obtained from Table 2), the average for the attribute selection and the 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, it is possible to see that the algorithms from Naïve Bayes and SVM presented inferior results. However, analyzing statistically the results obtained from this method with the original database results, we can observe that they are equivalent.

Table 17: General average of all databases (precision, in %).

Average	Naïve Bayes	C4.5	SVM	1-NN	3-NN	5-NN	7-NN
All attribute	75,6 ± 24,6	66,8 ± 13,1*	82,1 ± 21,4	68,2 ± 18,5	68,3 ± 23,3	69,3 ± 19,1	69,0 ± 17,3
Attribute Selection	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
Random Projection	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

4.4. Comparison between the original database and the dimensionality reduction methods

After obtaining the average of the results for each method, it is possible to compare all methods and the average results from the database, with all of the attributes as shown in Figure 4. Attributes Selection was the Dimensionality Reduction Method that showed the best results, followed by the combined use of Random Projection Method and Attributes Selection. Analyzing the results obtained from the Attributes Selection and Random Projection Method, it is possible to compare them. With the statistical analysis through hypothesis testing, it is concluded that the selection of attributes was better than the Random Projection Method. Analyzing the results from the selection of attributes and the approach proposed, it is possible to observe that the attributes selection results are higher in all classification algorithms. Statistical analysis proves this statement on the results of the algorithms Naïve Bayes, C4.5, and SVM, but the k-NN algorithm result (for k values used) is equivalent in both methods.

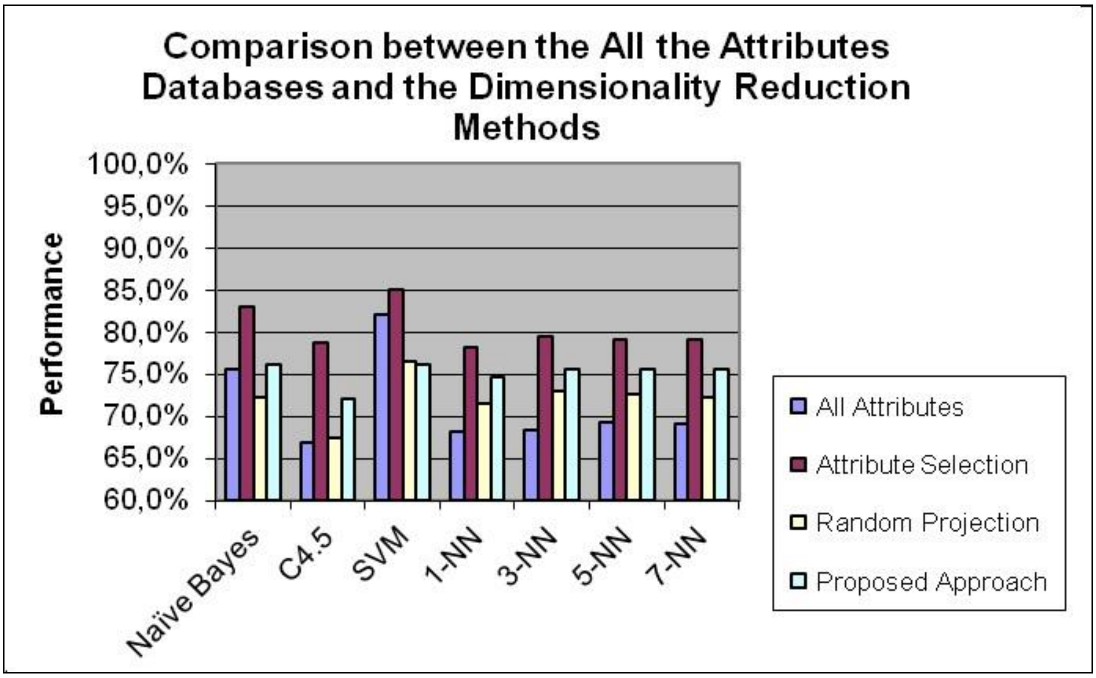


Figure 4: Comparative graph: All the Attributes Databases and Dimensionality Reduction Methods

Through this general comparison, it was possible to identify the best results more easily. It is observed that the results from the two Dimensionality Reduction Methods, Attributes Selection and the Random Projection Method, in three ways that were applied had better results compared to the results obtained when applied in databases with all the attributes. Analyzing the results, it is noted that the Attributes Selection was the Reduction Method which produced the best results. The use of the Random Projection Method contributed to the computational algorithms of time. However, the result obtained from this method was below the result from the Attribute Selection. The joint use of the Random Projection Method and the Attribute Selection had better results than the tests applied only in the Random Projection Method and on a smaller scale than the result of the Attribute Selection.

5. Conclusion

This work presented a new approach to Dimensionality Reduction by joining the transformation and selection of attributes. This approach was applied to gene expression databases. For the attributes selection, two main approaches were used: the Filter and the Wrapper. Analyzing the results obtained from these approaches, a significant improvement in the results was observed. When using the Attribute Selection algorithms, even in the worst cases, the classifier's hit rate was higher than the ones applied to databases with all the attributes. Also, there was a significant reduction in the number of attributes selected, mainly when the evaluation measures were applied with the sequential search. Comparing the results obtained from the two Attribute Selection Approaches (Filter and Wrapper), we observed the Wrapper Approach, when applied together with the sequential search, produced better results, followed by the measure of dependence evaluation belonging to the Filter Approach. The difference in the classifiers' results was small when looking at the hit rate, and the computational cost was much higher when the Wrapper Approach was used. In general, the execution of the algorithms belonging to the Filter Approach had a processing time in seconds to minutes. Nevertheless, the algorithms belonging to the Wrapper Approach had a processing time of the order of hours to days, which may, in some cases, become impracticable to its application. The SVM algorithm was generally the one that produced better results; however, its processing time was quite high compared to the other classification algorithms due to the size of the databases. Another algorithm that had good results in the classification was the Naïve Bayes. The Random Projection Method's application was an attempt to improve the execution time of the algorithms, especially the ones of Attributes Selection, and to increase even more the rate of adjustment of the classifier. When using the Random Projection Method, it could be observed a small increase in the hit rate in most of the classifiers compared to the results obtained when using the database with all the attributes. When the approach proposed was applied, it had a slightly greater improvement in the results when only the Random Projection Method was used. Although the result of applying the Random Projection Method, in the two cases that were used, was superior to the results obtained when no Reduction Method was used, it was not superior to the performance of the classification algorithms when only the Attribute Selection was used. Therefore, it is a general recommendation that Attributes Selection is a Dimensionality Reduction Method that yields great results when applied to gene expression bases. The Method of Random Projection is an alternative method since, besides reducing the computational cost when applied, mainly in conjunction with the selection of attributes, it produces good results. The experiments' results prove the applications of these Dimensionality Reduction Methods produce a higher classifier hit rate than when only the mining algorithm was applied to the databases with all the attributes. A new approach can be realized in future works. This approach is used in Stage 1, the Attributes Selection, and Stage 2, the Random Projection.

References

[1]. Ghodsi, A.: “Dimensionality reduction a short tutorial.” Department of Statistics and Actuarial Science, University of Waterloo, Ontario, Canada, 2006.

[2]. Maaten, L. V. D.; Postma, E., Herik, J. V. D.: “Dimensionality reduction: a comparative.” Journal of Machine Learning Research, vol. 10, pp. 66-71, 2009.

[3]. Almugren, N., Alshamlan, H.: “A Survey on Hybrid Feature Selection Methods in Microarray Gene

Expression Data for Cancer Classification." IEEE Access, vol. 7, pp. 78533-78548, Jun. 2019.

[4]. Witten, I. H., Eibe, F., Hall, M. A.: Data mining: practical machine learning tools and techniques. Burlington, MA: Morgan Kaufmann, 2016.

[5]. Manikandan, G., Abirami, S.: "A Survey on Feature Selection and Extraction Techniques for High-Dimensional Microarray Datasets" in Knowledge Computing and its Applications, Margret Anouncia S. and Wiil U. Ed. Singapore: Springer, 2018. pp. 311-333.

[6]. Liu, H., Motoda, H.: Feature Selection for Knowledge Discovery and Data Mining. New York: Kluwer academic Publishers, 1998, pp. 214.

[7]. Liu, H., Yu, L.: "Toward Integrating Feature Selection Algorithms for Classification and Clustering." IEEE Transactions on Knowledge and Data Engineering, vol. 17(4), pp. 491–502, 2005.

[8]. Dash, M., Liu, H.: "Feature selection for classification." Intelligent Data Analysis, vol. 1, pp. 131–156, 1997.

[9]. Hall, M.: "Correlation-based feature selection for discrete and numeric class machine learning." In Proc. 17th Proceedings of the International Conference on Machine Learning, 2000, pp. 359-366.

[10]. Kohavi, R.; John, G. H.: "The Wrapper Approach," in Feature Extraction, Construction and Selection: a data mining perspective, H. Liu and H. Motoda, Ed. New York: Springer US, 1998, pp. 33-49.

[11]. Cunningham, J. P., Ghahramani, Z.: "Linear dimensionality reduction: survey, insights, and generalizations." Journal of Machine Learning Research, vol. 16, pp. 2859-2900, 2015.

[12]. Jolliffe, I. T., Cadima, J.: "Principal component analysis: a review and recent developments." Philosophical Transactions of The Royal Society A Mathematical Physical and Engineering Sciences, vol. 374(2065), 2016.

[13]. Wang, L., Fu, X.: Data Mining with Computational Intelligence. Berlin: Springer, 2009.

[14]. Bingham, E., Mannila, H.: "Random projection in dimensionality reduction: applications to image and text data". In Proc. of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2001, pp. 245–250.

[15]. Bhui, N., Ram, P. K, Kuila, P.: "Feature Selection from Microarray Data based on Deep Learning Approach." In Proc. 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2020, pp. 1-5.

[16]. Borges, H. B., Nievola, J. C.: "Comparing the dimensionality reduction methods in gene expression databases." Expert Systems with Applications, vol. 39(12), pp. 10780-107958, 2012.

[17]. Kar, S., Das Sharma, K, Maitra, M.: "Gene selection from microarray gene expression data for classification of cancer subgroups employing PSO and adaptive K-nearest neighborhood technique." Expert Systems with Applications, vol.42(1), pp. 612-627, 2015.

[18]. Remeseiro, B., Bolon-Canedo, V.: "A review of feature selection methods in medical applications." Computers in Biology and Medicine, vol. 112, pp. 103375-103384, 2019.

[19]. Bertoni, A., Valentini, G.: "Random projections for assessing gene expression cluster stability." In Proc. 5th IEEE International Joint Conference on Neural Networks (IJCNN), 2005, pp. 149-154.

[20]. Khoirunnisa, A., Adiwijaya, Rohmawati, A. A.: "Implementing Principal Component Analysis and Multinomial Logit for Cancer Detection based on Microarray Data Classification." In Proc. 7th International Conference on Information and Communication Technology (ICoICT), 2019, pp. 1-6.

[21]. Cui, Y., Zheng, C., Yang, J.,Sha, W.: "Sparse maximum margin discriminant analysis for feature extraction and gene selection on gene expression data." Computers in biology and medicine, vol. 43(7), pp. 933-941, 2013.

[22]. You, W., Yang, Z., Yuan, M., Ji, G.: "Totalpls: local dimension reduction for multicategory microarray data." IEEE Transactions on Human-Machine Systems, vol. 44, pp. 125-138, 2013.

[23]. Sasikala, S., Appavu alias Balamurugan, S, Geetha, S. "Multi filtration feature selection (MFFS) to improve discriminatory ability in clinical data set." Applied Computing and Informatics, vol. 12(2), pp. 117-127, 2016.

[24]. Badaoui, F., Amar, A., Hassou, L. A., Zoglat, A., Okou, C.G.: "Dimensionality reduction and class prediction algorithm with application to microarray Big Data." Journal of Big Data, vol. 4, no. 32, Oct. 2017.

[25]. Elaziz, M. E. A.: "Simultaneous feature extraction and selection of microarray data using fuzzy-rough based multiobjective nonnegative matrix factorization." Journal of Intelligent & Fuzzy Systems, vol. 33(6), pp. 4043-4053, 2017.

[26]. Achlioptas, D.: "Database-friendly random projections." In Proc. 20th ACM SIGMODSIGACT-SIGART symposium on Principles of database systems, 2001, pp. 274–281.