Fusion Approaches of Feature Selection Algorithms for Classification Problems

Jhoseph Jesus, Daniel Araújo, and Anne Canuto

***Abstract*—The large amount of data produced by applications in recent years needs to be analyzed in order to extract valuable underlying information from them. Machine learning algorithms are useful tools to perform this task, but usually it is necessary to reduce complexity of data using feature selection algorithms. As usual, many algorithms were proposed to reduce dimension of data, each one with its own advantages and drawbacks. The variety of algorithms leads to either choose one algorithm or to combine several methods. The last option usually brings better performance. Based on this, this paper proposes an analysis of two distinct approaches of combining feature selection algorithms (decision and data fusion). This analysis was made in supervised classification context using real and synthetic datasets. Results showed that one proposed approach (decision fusion) has achieved the best results for the majority of datasets**

***Index Terms*—Feature Selection, Ensemble, Mutual Information, Data Analysis.**

I. Introduction

THE amount of data in the past few years has exponentially grown. Many of this data come from applications used to monitor events in the dynamic scenarios such as smart cities. The sets of existing data need to be analyzed and machine learning is one of the most suitable field to discover underlying relations inside the data and extract valuable information.

However, real world scenarios tend to have high complexity and, in order to build a more approximated model, a high number of variables (features) needs to be used. Problems in the fields of Bioinformatics, for instance, needs to have thousands of gene expressions measures to describe just a few dozens of patients [1]. Image processing, like segmentation or pattern discovery, uses pixels as features of images, resulting in huge number of features to describe one single image [2].

With such amount of features, most machine learning algorithms suffers in finding good solutions due to the curse of dimensionality. One suitable solution is to reduce the number of features since gathering more sample is often not possible [3].

To deal with this problem, several methods have been proposed in the past years. The general idea of reducing the dimensionality (number of features) of a dataset is to find a set of features that can represent the entire data in a way that the problem can be treated. This set can be composed of just a subset of the original data (feature selection) or can be a transformation of the initial features (feature extraction).

As mentioned before, a large number of algorithms were proposed to tackle this problem. They use distinct heuristics to find a solution and each one has its own domains, advantages and drawbacks. For example, Principal Component Analysis (PCA) [4], one of the most popular techniques, is based on linear projection of the largest eigenvector of the correlation matrix to the original features, which means that it is very sensitive to the magnitude of values and, by consequence, to simple rotations and/or translation in data [3].

Recently, Information Theory descriptors, initially used to measure the efficiency of data transmission [5], are been used to quantify information in a variety of real world problems. The Dimension Reduction (DR) problem is one of them. For instance, [6] proposed a series of Mutual Information based techniques to select the most relevant features of a dataset regarding to the given classes of the problem. Algorithms based on Mutual Information could be a better choice than traditional linear methods because they can actually measure the dependency of two variables, including non-linear correlation, which is very common in real world situations.

Unfortunately, there is no better algorithm to treat all problems. So, in order to reduce a dimensionality of one dataset, a researcher has either to know very well the DR algorithms and the data to choose the best possible method or he has to randomly choose one expecting that it can perform the task well enough.

However, one common alternative used by ML researches can be used in the context of feature selection: the combination or ensemble approach. This approach often used several methods and combines their outputs to produce one single solution, probably better than the single ones.

Aiming at contributing for this important subject, this paper aims to analyze two distinct apporaches of combining multiple feature selection algorithms. The first, combines solutions produced by different feature selection algorithms using a voting scheme to create a single solution. The idea of this approach is to have a combination of data, obtained by the different feature selection algorithms (data fusion). The second one is based on an ensemble of classification algorithms trained by datasets reduced by feature selection algorithms (decision fusion).

The rest of this paper is organized as follows: Section II brings the details of the two approaches used in this paper to combine solutions produced by feature selection algorithms. Section III shows the algorithms and datasets used to perform the experiments. Finally, Section IV presents the results and

J. Jesus and D. Araújo are with the Digital Metropolis Institute, Federal University of Rio Grande do Norte, Natal/RN, Brazil, e-mail: jhoseph.kelvin@gmail.com, daniel@imd.ufrn.br

A. Canuto is with the Department of Informatics and Applied Mathemathics, Federal University of Rio Grande do Norte, Natal/RN, Brazil, e-mail: anne@dimap.ufrn.br

J. Jesus is also with the Department of Informatics and Applied Mathemathics, Federal University of Rio Grande do Norte

brings a brief discussion closed by conclusions in Section V.

II. The proposed Fusion Approaches

This section will bring information about the overall operation of the proposed combination approaches to be used in this paper, data and decision fusion.

A. Data Fusion

One simple way of combining features obtained by several feature selection algorithms is to use a voting scheme to choose the most relevant features, based on the output of different feature selection algorithms. In other words, this approach provides a fusion of features selected by different feature selection algorithms and uses a voting strategy to select the most important ones. The voting technique is based on the relevance that features appears in the outputs of each algorithm. Combining feature selection algorithms has been successfully used in the pattern recognition literature, such as in [7] and in [8]. The last work brings a similar study about combination approaches, with other fusions strategies. A general overview of the data fusion approach can be seen in Figure 1.

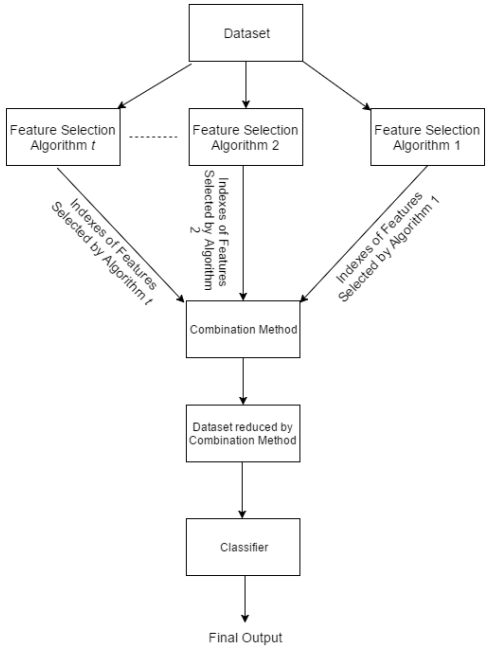

Fig. 1. Data Fusion.

Let $X_{n\times m}$ be a dataset and $S_{n\times k}$ be the reduced dataset, where n represent the number of instances, m the original number of attributes and $k < m$ the subset of selected features. In order to have a combined solution, we have to run t algorithms, where $t > 1$, and each algorithm selects from X a subset of features $\mathbf{f} = f_1, f_2, ..., f_l$, where $l \leq m$. One can reduce a dataset using the output of the FS algorithm to extract a subset of X, $S'_{n\times l}$. It is important to notice that the general results from feature selection algorithms are indexes of the features to be selected.

Using the selected features **f**, we just count how many times each feature appears in each solution found by the FS algorithms weighted by their relevance. The relevance, in this context, is inversely proportional to its position in the feature vector. So, the relevance of the f_i feature can be defined by:

$$r_i = \frac{1}{j} \tag{1}$$

where j is the position of the feature in the output vector. For example, if a feature is the first choice of an algorithm, its relevance is equals to one. If it appears in the fourth position, then its relevance is 0.25. Using this strategy we consider not only the presence of a feature in the DR output, but its importance to the whole process.

Now, consider $F = \{\mathbf{f}^1, \mathbf{f}^2, ..., \mathbf{f}^t\}$ as the set of all outputs created by t algorithms. We can define a voting factor for each feature as:

$$v_i = \sum_{j=1}^{t} r_i. \tag{2}$$

where $v = \{v_1, v_2, ..., v_l\}$ is the set of voting factors for each feature. In other words, we basically sums up the relevance of the features for each algorithm. After this step, we just select the k features with highest values in v vector to have the more relevant features for all t algorithms and build the S dataset.

B. Decision Fusion

This section presents a second approach to combine multiple feature selection algorithms. The idea is to use the structure of ensemble of classifiers as a fusion approach, in which the decision of the classification algorithms will be combined in the combination method of the ensemble system (Decision fusion). The concept of ensemble systems has emerged in the last decades as a strategy for combining classifiers, aiming to provide a solution that is potentially more efficient than any single component [9]. An ensemble system consists of a set of c individual classifiers (ICs) that are organized in a parallel way. The set of ICs receives the input data and their outputs are sent to the combination module ($Comb$) that provides the overall answer of the ensemble. Therefore, unlabelled patterns $\{U_i \in R^d | i = 1, 2, ..., n\}$ will be presented to all individual classifiers and a combination method combine their output to produce the overall output of the system $O = Comb(y_j), \{y_j = (y_{j1}, ..., y_{jk} | j = 1, ..., c \text{ and } k = 1, ..., r\}$, where the number of individual classifiers is defined by c and r describes the number of labels in a dataset. For ensemble systems, the main aim is that individual components offer complementary information about an input pattern and this complementary information tends to increase the effectiveness of the whole recognition process [10].

In this context, the idea consists of combining feature selection algorithms using a homogeneous ensemble of classification algorithms. That is, we do not combine the outputs created by feature selection algorithms but decisions provided by the classification algorithms trained with the reduced datasets produced by feature selection algorithms. The general idea of the ensemble approach can be seen in Figure 2.

As we can see, we first use feature selection algorithms to produce distinct subsets of the original X dataset. Those

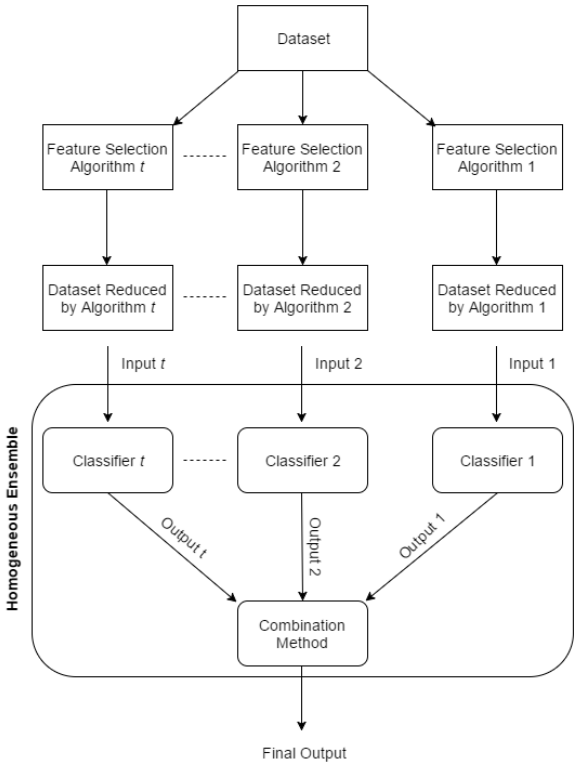

Fig. 2. Decision Fusion.

subsets are used as training data to classification algorithms that are latter combined to have on single solution. It is important to notice that the number of individual classifiers c is set by the number of feature selection algorithms.

The main goal of using a method based on ensemble classifiers in the context of this paper is to increase the diversity during the process of feature selection opposing the combination method which has low diversity.

III. Material and Methods

A. Dimensionality Reduction Algorithms

One of the main purposes of this paper is to present two approaches to combine solutions provided by different feature selection algorithms (dimensionality reduction algorithms). Mutual Information based algorithms have a high potential to perform feature selection specially when compared to more traditional methods. Therefore, all feature selection algorithms to be used in this paper are based on Mutual Information.

In order to validate our fusion approaches, we followed the approach used in [11] and in [12]. We selected four Mutual Information based algorithms from the first paper and one from the second, including the author's algorithm, Spectral relaxation global Conditional Mutual Information (SPEC_CMI). Even that all the five algorithms use the same heuristic, each one of them tries to reach different goals. They use the same information about the data to answer different questions which leads us to have more diversity in our experiment. The following topics will describe the algorithms and their respective objectives. As the authors turned public their Matlab®toolbox implementations [1] [2] we have used it to run our experiments. The following algorithms were used (more details can be found in cited references):

- Quadratic programming feature selection (QPFS) [13]: QPFS algorithm selects features reducing the task of selection to a quadratic optimization problem, using the Nystrom method for approximate matrix diagonalization, which gives to this method the capacity of dealing with very large datasets. This solution represents a faster way to select features, when compared to other methods of feature selection with mutual information.
- Spectral Relaxation Global Conditional Mutual Information (SPEC_CMI) [11]: This algorithm implements a systematic approach to the problem of global Mutual Information (MI)-based feature selection via spectral relaxation techniques. This approach treats issues commonly faced by other greed algorithms, like feature 'self-redundancy' and thus not leading to sub-optimal solutions.
- Maximum Relevance Minimum Total Redundancy (MRMTR) [14]: MRMTR algorithm selects features using minimal-redundancy and maximal-relevance heuristics. The criterion function represented by a multi-objective problem, aims to select a set of features which jointly have larger relevance on the target class and at the same time have less redundancy between them.
- Conditional Mutual Information Maximization (CMIM) [15]: CMIM algorithm is based on conditional mutual information, that measures the amount of mutual information of two random variables with respect to a third. The algorithm works picking features that maximize the mutual information of selected features with the class to predict conditional to any feature already selected, ensuring the selection of features are both individually informative and weakly dependent.
- Mutual Information Feature Selection (MIFS) [16]: MIFS algorithm is based on greedy selection of features and considers both mutual information with respect to the output class and with respect to the already selected features. Given a set of features, the algorithm chooses as the next feature the one that maximizes the information about the classes. That is, a feature must be informative about the class without being predictable from the current set of features.

All feature selection algorithms were set with their default parameters. More details about the previous algorithms can be found in [11] and [12]. The two combination methods were used to ensemble all the solutions to produce a single output. We used two ways to combine the features, the first is a combination algorithm showed in Section II-A and the second is an approach based on ensemble of classifiers described in Section II-B.

In addition, for sake of performance comparison, we will also compare the results obtained by the fusion approaches

[1] avaliable at http://www.mathworks.com/matlabcentral/fileexchange/47129-information-theoretic-feature-selection

[2] avaliable at http://www.mathworks.com/matlabcentral/fileexchange/26981-feature-selection-based-on-interaction-information

to a widely used dimensionality reduction algorithm, PCA (Principal Component Analysis) [17], [18].

In order to compare the two approaches analyzed in this paper, we had to set the number of dimensions of the reduced dataset (target dimension). We selected to use three features for each feature selection method. Two main reasons are decisive in this selection, which are: we can test the power of feature selection algorithms to resume the underlying information using few data; and three dimensions is a common configuration used to visualize data.

B. Classification Algorithms

In order to evaluate the performance of the proposed approaches, We used three classification algorithms, which are: Decision Tree [19], Naive Bayes [19] and k Nearest Neighbor (k-NN) [3]. They have been chosen because they are widely used in the machine learning community and each one has a distinct approach to find the best solution. Based on this, we tried to cover a wide range of heuristics to classification and to avoid a possible bias to a specific approach.

In order to run all algorithms, we used two tools: the Weka software [20] and Matlab software [21] with all parameters set to default. We are aware that a fine tuning of parameters would probably lead to better results, but the number of variables handled in the experiments was already too high. As the main purpose of this paper is to analyze the feature selection algorithms and the two approaches to combine solutions, giving the same settings to all should be enough.

With the purpose to achieve more robust results, we used a 10-fold-cross-validation approach for both combination and ensemble methods proposed. In addition, all algorithms were executed 10 times and we computed the average results and respective standard deviations.

In order to compare the effectiveness of the fusion approaches, a statistical test was applied, which is called hypothesis test (one-tailed student t-test), with a confidence level of 95% (= 0.05).

C. Datasets

In our experiments, we used ten datasets from distinct natures, distributed in two sets, which are called: artificial and real. The artificial datasets are composed of three datasets: `Gaussian`, `Simulated` and `Friedman`. The real datasets are composed of seven datasets: `LSVT`, `Lung Cancer`, `Breast Cancer Diagnostic`, `Connectionist Bench`, `Ionosphere`, `Jude` and `Colon Cancer`.

With the exception of `Lung Cancer` [1], `St Jude Leukemia` [22], and `Colon Cancer` [23] which were collected at the Bioinformatics Research Group of Seville repository [24], all datasets were collected at the UCI machine learning repository [25]. Those datasets were also used in some papers with similar purposes.

Datasets were selected aiming to cover different ranges of number of samples and features. The main characteristics of each dataset are presented in Table I, where n is the number of samples, C is the number of classes and d is the number of features (dimensionality).

Dataset	n	C	**Dist.of Classes**	d
`LSVT`	126	2	42,84	310
`Lung Cancer`	181	2	31,150	12533
`Breast Cancer Diagnostic`	569	2	212,357	30
`Connectionist Bench`	208	2	97,111	60
`Ionosphere`	351	2	126,225	32
`St Jude Leukemia`	248	6	15,27,64,20,40,79	985
`Gaussian`	60	3	20,20,20	600
`Simulated`	60	5	8,12,10,15,5,10	600
`Friedman`	1000	2	436,564	100
`Colon Cancer`	62	2	22,40	2000

TABLE I
DATASETS DESCRIPTON.

`LSVT` dataset [26] is composed of 126 sustained vowel /a/ phonations features with 310 dysphonia measures aiming to do a characterization of speech signals of Parkinson Disease subjects.

`Lung Cancer (LungC)` is a gene expression dataset used in [1] to study malignant pleural mesothelioma (MPM) and adenocarcinoma (ADCA) of the lung. There are 181 tissue samples (31 MPM and 150 ADCA), each sample is described by 12533 genes.

`Breast Cancer Diagnostic (BreastC)` dataset [27] is composed of 569 patterns of cells nuclei computed from the digitization image of a needle aspirate of a breast mass. Those images features are describe by 32 attributes.

`Connectionist Bench (ConnB)` dataset [28] was created from 208 patterns of sonar signals that bounced off metal cylinders (111 samples) or rocks (97 samples) in several distinct angles.

The `Ionosphere (Iono)` dataset [29] is composed of 351 radar returns from the ionosphere divided as either suitable for further analysis or not.

`St Jude Leukemia (Jude)` dataset [22], [30] was generated from gene expression data of leukemia cells. There are 248 samples of leukemia cells and 985 genes as attributes describing the expression level of each gene to each sample of cell.

`Gaussian (Gauss)` and `Simulated (Simul)` datasets are synthetic databases that simulate microarray data and were created to test the ML algorithms in the gene expression analysis [22]. Both datasets have 60 samples and 600 attributes.

`Friedman (Fried)` dataset is a artificially dataset originated by the Friedman's funcion from [31]. Friedman's function is used to generate data, including both linear and non-linear relations between samples and output, and a normalized noise added to the output. The `Friedman` dataset is composed of 1000 samples and 100 attributes.

`Colon Cancer (ColonC)` dataset [23] represent a set of broad patterns of gene expression revealed by clustering of tumor and normal colon tissues probed by oligonucleotide arrays, being composed of 2000 genes as attributes and 62 tissues samples (22 normal colon tissues and 40 tumor colon tissues).

IV. Results and Discussion

In this section, we will present the results of the empirical analysis performed with all five approaches, two fusion approaches, PCA and the original dataset (no feature selection).

In addition, we use three different classification algorithms and the next three subsections will describe the obtained results for these algorithms.

A. Decision Trees

Table II presents the accuracy and standard deviation of the decision tree using all four feature selection approaches, as well as the original dataset, for all ten datasets. For each dataset (each line of the table), the bold number represents the approach with best performance (highest accuracy). In addition, we applied the statistical test and (*) represents the cases in which there is statistical difference in performance, in relation to the approach with best performance.

CA	Decision Tree			
DRM	DecFusion	DatFusion	PCA	Original
DRA	Mean±Std	Mean±Std	Mean±Std	Mean±Std
LSVT	83.75±2.11	**85.94±8.94**	75.05±9.67(∗)	74.49±13.06(∗)
LungC	**98.24±0.51**	94.81±4.06(∗)	92.22±5.36(∗)	93.99±5.00(∗)
BreastC	**94.65±0.53**	93.36±3.11	94.31±2.91	93.27±3.55
ConnB	**76.39±1.83**	72.56±9.47(∗)	74.27±10.45	73.61±9.34
Iono	**91.85±0.66**	90.11±4.64	85.84±5.05(∗)	91.06±4.16
Jude	84.63±1.08(∗)	83.07±7.08(∗)	**93.87±4.44**	89.31±5.67
Gauss	73.18±3.54(∗)	75.50±16.99(∗)	**98.00±6.39**	52.33±19.39(∗)
Simul	68.00±3.01(∗)	68.50±15.51(∗)	**86.83±13.46**	74.83±15.26(∗)
Fried	88.18±0.70	**89.42±2.95**	68.71±4.63(∗)	86.24±3.54
ColonC	**86.74±2.43**	83.79±12.92	61.92±9.97(∗)	81.95±13.86

TABLE II
RESULTS USING DECISION TREE CLASSIFIER.

As it can be observed from Table II, one proposed approach provided the best performance for almost all datasets. The only exception were the St Jude, Gaussian and simulated datasets. The decision fusion provided the best performance in five datasets and data fusion provided the best performance in three datasets. The statistical test showed that the decision fusion approach had statically better performance than PCA in 3 datasets and than the original data in only 1 dataset. In the same sense, the statistical test showed that the data fusion approach had statically better performance than PCA in 2 datasets and than the original data in 1 dataset.

B. Naive Bayes

Table III presents the accuracy and standard deviation of the Naive Bayes algorithm using all four feature selection approaches, as well as the original dataset, for all ten datasets.

CA	Naive Bayes			
DRM	DecFusion	DatFusion	PCA	Original
DRA	Mean±Std	Mean±Std	Mean±Std	Mean±Std
LSVT	**72.79±0.79**	69.64±11.64	54.88±12.04(∗)	54.40±12.59(∗)
LungC	**99.46±0.08**	98.29±3.30	91.06±5.57(∗)	97.95±3.21(∗)
Breastc	**96.01±0.13**	94.24±3.30(∗)	91.60±3.63(∗)	93.30±3.30(∗)
ConnB	**76.63±0.73**	68.54±10.32(∗)	76.18±9.04	67.71±8.66(∗)
Iono	86.74±0.20	**89.29±4.64**	83.39±6.09(∗)	81.54±6.10(∗)
Jude	88.59±0.44(∗)	85.00±6.26(∗)	95.12±3.77	**98.22±2.39**
Gauss	80.10±2.92(∗)	67.67±18.32(∗)	**100.00±0.00**	79.66±17.82(∗)
Simul	73.00±1.69(∗)	67.50±15.24(∗)	87.16±9.43(∗)	**91.66±8.3**
Fried	63.12±0.31	63.05±4.14	59.59±3.80	**64.83±4.52**
ColonC	**80.42±1.59**	78.83±16.04	56.02±18.92(∗)	55.69±17.04(∗)

TABLE III
RESULTS USING NAIVE BAYES CLASSIFIER.

As it can be observed from Table III, the proposed approaches provided the best performance half of the analysed datasets. The decision fusion provided the best performance in four datasets and data fusion provided the best performance in only one dataset. The result of the statistical test showed that the increase in performance of the decision fusion approach was proved to be statistically significant in 4 datasets, in comparison to PCA, and in all 5 datasets, in comparison to the original data. In the same sense, the statistical test showed that the data fusion approach had statically better performance than PCA and original data in the `Iono` dataset.

When comparing the performance of both proposed approaches with NB, in relation to the results obtained in the previous section, the proposed approaches obtained the best peformance in less datasets (7 for DT and 5 fo NB). However, these improvements proved to be statistically significant more frequently (7 for DT and 11 for NB).

C. k Nearest Neibourgh

Table IV presents the accuracy and standard deviation of the *k*-NN (nearest neighbourh) algorithm using all four feature selection approaches, as well as the original dataset, for all ten datasets.

CA	k-NN			
DRM	DecFusion	DatFusion	PCA	Original
DRA	Mean±Std	Mean±Std	Mean±Std	Mean±Std
LSVT	**84.80±9.20**	75.75±1.61(∗)	73.10±12.21(∗)	75.88±12.49(∗)
LungC	**99.01±0.29**	98.73±2.71	95.47±4.46(∗)	95.19±4.14(∗)
BreastC	90.49±0.41(∗)	**95.82±2.83**	92.64±3.48(∗)	95.64±2.32
ConnB	75.51±1.06(∗)	67.48±10.77(∗)	71.63±9.67(∗)	**86.17±8.45**
Iono	**92.99±0.46**	88.38±5.06(∗)	85.86±5.48(∗)	87.16±4.96(∗)
Jude	86.72±0.80(∗)	80.47±6.66(∗)	92.43±4.31(∗)	**98.67±2.07**
Gauss	75.87±2.71(∗)	59.33±18.09(∗)	**100.00±0.00**	98.33±5.03
Simul	79.60±1.64	52.17±16.86	87.16±10.56(∗)	**100.00±0.00**
Fried	**90.84±0.37**	90.57±2.92	61.62±4.43(∗)	52.96±4.55(∗)
ColonC	**84.27±1.20**	79.62±14.02(∗)	62.73±17.71(∗)	76.83±17.21(∗)

TABLE IV
RESULTS USING K-NN CLASSIFIER.

As it can be observed from Table IV, the decision fusion provided the best performance in half of the datasets, while the data fusion approach provided the best performance in only one dataset. The results statistical test showed that the decision fusion approach had statically better performance than PCA and the original dataset in all 5 datasets. However, for the data fusion approach, the statistical test showed that the this approach had statically better performance only for PCA (`BreastC` dataset), and similar performance than the original data. The results obtained by the *k*-NN algorithm are very similar to the ones obtained by the NB algorithms.

In summary, based on the empirical analysis conducted in this paper, we could state that the decision fusion approach is the best feature selection method, comparing to the data fusion approach and PCA, providing better performance in the majority of datasets. In addition, it could improve performance, when compared to the original data (no feature selection), in the majority of datasets.

V. CONCLUSIONS

This paper presented two distinct apporaches of combining multiple feature selection algorithms. The first one combines solutions produced by different feature selection algorithms

using a voting scheme to create a single solution. The idea of this approach is to have a combination of data, obtained by the different feature selection algorithms (data fusion). The second one is based on an ensemble of classification algorithms trained by datasets reduced by feature selection algorithms (decision fusion).

In order to assess the performance of the proposed approaches, an empirical analysis was conducted. In this analysis, the proposed approach used three different classification algorithms (DT, NB and *k*-NN) and they were all applied to 10 different datasets. For comparison purposes, we also applied a standard PCA algorithm and the original data (no feature selection).

Through this analysis, we could state that the decision fusion approach is the best feature selection method, comparing to the data fusion approach and PCA, providing better performance in the majority of datasets. In addition, it could improve performance, when compared to the original data (no feature selection), in the majority of datasets.

As future work we can investigate other feature selection techniques. The results can be complemented by performing experiments with heterogeneous ensembles and other classification algorithms.

ACKNOWLEDGMENT

This paper was partially supported by CNPq Universal Grant no 480997/2013-6 and UFRN scholarship program.

REFERENCES

[1] G. J. Gordon, R. V. Jensen, L. li Hsiao, S. R. Gullans, J. E. Blumenstock, S. Ramaswamy, W. G. Richards, D. J. Sugarbaker, and R. Bueno, "Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma," *Cancer Res*, vol. 62, pp. 4963–4967, 2002.

[2] R. C. Gonzalez and R. E. Woods, *Digital Image Processing (3rd Edition)*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 2006.

[3] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1988.

[4] I. Jolliffe, *Principal Component Analysis*, ser. Springer Series in Statistics. Springer, 2002. [Online]. Available: https://books.google.com.br/books?id=_olByCrhjwIC

[5] C. E. Shannon, "A mathematical theory of communication," *Bell system technical journal*, vol. 27, 1948.

[6] G. Brown, A. Pocock, M.-J. Zhao, and M. Luján, "Conditional likelihood maximisation: A unifying framework for information theoretic feature selection," *J. Mach. Learn. Res.*, vol. 13, pp. 27–66, Jan. 2012. [Online]. Available: http://dl.acm.org/citation.cfm?id=2188385.2188387

[7] Q. Shen, R. Diao, and P. Su, "Feature selection ensemble," in *Turing-100. The Alan Turing Centenary*, ser. EPiC Series in Computing, A. Voronkov, Ed., vol. 10. EasyChair, 2012, pp. 289–306.

[8] R. C. Prati, "Combining feature ranking algorithms through rank aggregation," in *The 2012 International Joint Conference on Neural Networks (IJCNN), Brisbane, Australia, June 10-15, 2012*, 2012, pp. 1–8. [Online]. Available: http://dx.doi.org/10.1109/IJCNN.2012.6252467

[9] R. R. Parente, A. M. P. Canuto, and J. C. X. Jr., "Characterization measures of ensemble systems using a meta-learning approach," in *The 2013 International Joint Conference on Neural Networks, IJCNN 2013, Dallas, TX, USA, August 4-9, 2013*, 2013, pp. 1–8. [Online]. Available: http://dx.doi.org/10.1109/IJCNN.2013.6707016

[10] L. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms*. New York: Wiley, 2004.

[11] X. V. Nguyen, J. Chan, S. Romano, and J. Bailey, "Effective global approaches for mutual information based feature selection," in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '14. New York, NY, USA: ACM, 2014, pp. 512–521. [Online]. Available: http://doi.acm.org/10.1145/2623330.2623611

[12] G. Brown, "A new perspective for information theoretic feature selection," in *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics (AISTATS-09)*, D. V. Dyk and M. Welling, Eds., vol. 5. Journal of Machine Learning Research - Proceedings Track, 2009, pp. 49–56. [Online]. Available: http://jmlr.csail.mit.edu/proceedings/papers/v5/brown09a/brown09a.pdf

[13] I. Rodriguez-Lujan, R. Huerta, C. Elkan, and C. S. Cruz, "Quadratic programming feature selection," *J. Mach. Learn. Res.*, vol. 11, pp. 1491–1516, Aug. 2010. [Online]. Available: http://dl.acm.org/citation.cfm?id=1756006.1859900

[14] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2005.159

[15] F. Fleuret, "Fast binary feature selection with conditional mutual information," *J. Mach. Learn. Res.*, vol. 5, pp. 1531–1555, Dec. 2004. [Online]. Available: http://dl.acm.org/citation.cfm?id=1005332.1044711

[16] R. Battiti, "Using mutual information for selecting features in supervised neural net learning," *Trans. Neur. Netw.*, vol. 5, no. 4, pp. 537–550, Jul. 1994. [Online]. Available: http://dx.doi.org/10.1109/72.298224

[17] J. Jackson, *A User's Guide to Principal Components*, ser. Wiley Series in Probability and Statistics. Wiley, 2005. [Online]. Available: https://books.google.co.in/books?id=f9s6g6cmUTUC

[18] I. T. Jolliffe, *Principal Component Analysis*. Berlin; New York: Springer-Verlag, 1986.

[19] T. M. Mitchell, *Machine Learning*, 1st ed. New York, NY, USA: McGraw-Hill, Inc., 1997.

[20] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: An update," *SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 10–18, Nov. 2009. [Online]. Available: http://doi.acm.org/10.1145/1656274.1656278

[21] MATLAB, *version 7.10.0 (R2010a)*. Natick, Massachusetts: The MathWorks Inc., 2010.

[22] S. Monti, P. Tamayo, J. Mesirov, and T. Golub, "Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data," *Machine Learning*, vol. 52, no. 1, pp. 91–118, 2003. [Online]. Available: http://dx.doi.org/10.1023/A:1023949509487

[23] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 96, no. 12, pp. 6745–6750, Jun 1999.

[24] B. G. Seville, "BIGS bioinformatics research group of seville repository," 2004.

[25] M. Lichman, "UCI machine learning repository," 2013. [Online]. Available: http://archive.ics.uci.edu/ml

[26] A. Tsanas, M. A. Little, C. Fox, and L. O. Ramig, "Objective automatic assessment of rehabilitative speech treatment in parkinson's disease," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 22, no. 1, pp. 181–190, Jan 2014.

[27] N. Street, W. H. Wolberg, and O. L. Mangasarian, "Nuclear feature extraction for breast tumor diagnosis," 1993.

[28] P. R. Gorman and T. J. Sejnowski, "Analysis of hidden units in a layered network trained to classify sonar targets," *Neural Networks*, vol. 1, no. 1, pp. 75–89, 1988.

[29] V. G. Sigillito, S. P. Wing, L. V. Hutton, and K. B. Baker, "Classification of radar returns from the ionosphere using neural networks," *Johns Hopkins APL Technical Digest*, vol. 10, pp. 262–266, 1989.

[30] E. J. Yeoh, M. E. Ross, S. A. Shurtleff, W. K. Williams, D. Patel, R. Mahfouz, F. G. Behm, S. C. Raimondi, M. V. Relling, A. Patel, C. Cheng, D. Campana, D. Wilkins, X. Zhou, J. Li, H. Liu, C. H. Pui, W. E. Evans, C. Naeve, L. Wong, and J. R. Downing, "Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling," *Cancer Cell*, vol. 1, no. 2, pp. 133–143, Mar 2002.

[31] J. H. Friedman, "Stochastic gradient boosting," *Comput. Stat. Data Anal.*, vol. 38, no. 4, pp. 367–378, Feb. 2002. [Online]. Available: http://dx.doi.org/10.1016/S0167-9473(01)00065-2