International Journal of Computer Information Systems and Industrial Management Applications.
ISSN 2150-7988 Volume 8 (2016) pp. 214–225

Improving the stability of wrapper variable selection applied to binary classification

Silvia Cateni and Valentina Colla

TeCIP Institute,
Scuola Superiore Sant' Anna,
Pisa, Italy
s.cateni{colla}@sssup.it

Abstract: **Wrapper variable selection methods are widely adopted in many applications, among which the design of classifiers. The main problem related to these approaches regards the stability of the selection, namely the exploitation of different training data set can lead to the selection of different variable subsets. This problem is particularly critical in applications where variable selection is used to interpret the behaviour of the process or phenomenon under consideration, i.e. to understand which among a potentially huge list of variables actually affect the classification. The paper proposes a method that improves the stability of the wrapper variable selection procedures while preserving and possibly improving the classification performance. Moreover three binary classifiers are performed in order to prove the effectiveness of the proposed method.**
Keywords: Variable selection, Wrapper, Binary classification, Data mining, Stability

I. Introduction

Pre-processing of the data is essential to efficiently exploit machine learning techniques. Variable selection is a fundamental preliminary step concerning the analysis of the data to be exploited for the design of a large variety of models or systems which are based on a self-learning or training procedure exploiting experimental data. Variable selection is the process of selecting a subset of relevant variables in a list of measured features which can affect a given system or phenomenon in order to use them to build a model representing the phenomenon itself [1]. Variable selection is therefore crucial in a wide list of domains, such as machine learning [2] [15], pattern recognition [3] and data mining [4]. Moreover variable selection has been widely performed in different applications: function approximation [5] [6], classification [7] [8] [9] and clustering [10]. The importance of variable selection, which has been widely studied for a long time now [11] [4] [12], has been enhanced also by the always increasing growth and development of sensing tools and data storage capabilities which are available in real word applications [13] [14], such as industries and public services, that provide access to huge amount of different data that, on one hand, allow the development of more complex and reliable model but, on the other hand, provide ever more challenging task from the knowledge extraction point of view [15].
When designing any form of statistical or Artificial Intelligence (AI) based classifier, especially related to a phenomenon or system which is poorly or only partially known, the very first problem to address is the selection of the correct input variables for the system itself. It is known that an appropriate subset of input variables could provide better performances than the whole set [16]. This phenomenon, often called *peak effect* can be explained by considering that building a true-minimum classification error from a finite training set is impossible and the approximation is affected by irrelevant features. The selection of a suitable set of input variables is computationally advantageous and can also improve the classification accuracy [11].
Although it is relevant to compute the performance of a variable selection algorithm, it is worth noting that by removing irrelevant and redundant features the generalization capability increases and the model interpretability improves. Thus a benefit can also be gained on the acquisition of knowledge on the considered process or phenomenon, as it is possible to understand which factors mostly affect it. Nonetheless, beside high performance and computational efficiency, *stability* is a crucial factor for evaluating feature selection algorithms [17]. The stability of a variables subset is defined as the sensitivity of a classification method when this variables set is used as input with respect to variations in the training set. Stability is also crucial when the aim is knowledge discovery and not only an accurate classification. In fact a good feature selection algorithm should not only improve the classifier performance but also provide stable selection results when the training data sets are modified.
In particular, in the present paper the design of binary classifiers is addressed, which has a relevant importance from the practical point of view, as many real world any real-world applications related e.g. to anomalies detection and forecasting are approached as binary classification problems.
For instance intrusion detection, which plays an important role in the protection of communication networks, is often formalised as a binary classification problem faced by pattern recognition systems, whose performance is highly dependent on the features which are used as inputs. To this aim in [18] an approach for the selection of optimal feature subset

based on the analysis of the Pearson correlation coefficients is discussed, which is capable to increase the performance of classifiers applied to distinguish whether a considered system activity is "intrusive" or "legitimate". Recent comparative study on the application of variable selection to intrusion detection by means of binary classifiers can be found in [19] and [20].
In the medical field, some diseases diagnosys problems are approached through binary classifiers [21] [22] [23]. Another exemplar application of binary classification which can benefit from a preliminary feature selection stage is the analysis of structural and functional data related to the humane genome. Its relationship with particular diseases has represented a challenge for the data mining community [24], leading to the development of ad-hoc procedures and algorithms. Also in this application the binary classification is very used such as, for instance, in [25], [26] and [27].
Finally in the industrial field faults and anomalies diagnosys or forecasting and defective products identification in quality monitoring in the industrial field are often faced through binary classifiers. The progress of ICT and the developments of the sensing technologies allows to equip industrial plants with an ever increasing number of sensors which collect a huge amount of information. However, especially for large and very complex processes where a series of chemical, physical and thermo-mechanical reactions simultaneously occurr (e.g. in process industry), it is difficult to identify the factors which mostly affect faults and quality problems. Variable selection can be applied to this purpose [7] [9], as a mean to increase the knowledge of the phenomenon and to select the sensorial information which is mostly relevant for faults monitoring and forecasting.
The present paper addresses the variable selection problem with a particular attention to improve the stability of the algorithm for the selection of variables subset applied to binary classifier. To this aim the most frequent couples of variables are considered, in order to take into account the mutual interaction between them and not only the contribute of one variable individually exploited.
The paper is organized as follows: Sec. II provides a brief overview on the variable selection techniques; Sec. III describes stability problem. Sec. IV is focused on the description of the proposed approach; Sec. V describes the developed experiments and discusses the obtained numerical results. Finally Sec. VI provides some concluding remarks.
This paper is an extended version of the paper entitled 'Improving of the stability of sequential forward variables selection'which was presented by the same authors at the 15th International Conference on Intelligent Systems Design and Applications ISDA 2015 [28].

II. Background on the Variable Selection

From a generic perspective, given any kind of self-learning system which needs to produce any form of output (e.g. the predicted or estimated value of one or more variables or a binary classification output) on the basis of the values of some input features belonging to a possibly very large set of potentially relevant variables, variable selection performs a reduction of the dimension of the features set and highlights those ones that mostly affect a given phenomenon. The main objectives of variable selection include: data dimensionality reduction, model accuracy improvement and achievement of a deeper knowledge and a more accurate representation of the considered phenomenon [29].
Variable selection methods can be categorized into three main classes: filter, wrapper and embedded approaches.
Filters select the best subset of input variables independently from the adopted learning algorithm [30]. The subset is created by considering the relationship between input and output variables of the designed system and therefore all variables are classified on the basis of their pertinence to the target by performing a statistical test [31] [32] [33]. The main advantage of filters lies in their low computational complexity, which makes them fast; their main disadvantage resides in their inability to optimize the model adopted in the learning machine [34]. A simple example of filters is represented by the correlation-based approach which calculates the correlation coefficient between each variable and the target, features are then ranked and a subset is extracted containing the variables with the highest correlation coefficient. This method is very fast and the removal of features with a low correlation coefficient lowers the redundancy of the input set. However the linear correlation approach is inadequate when dealing with real-world datasets, where variables are often correlated in a highly nonlinear way [35]. Other widely applied filter approaches are the chi-square approach [36] and the Information Gain method [37].
Wrappers consider the machine learning as a black box and select the optimal subset of features by considering their predictive power. Their main advantage lies in the selection of the optimal subset considering the performance of the machine learning algorithm. Moreover, being designed as a black box system, wrappers approaches are simple and universally applicable. Two further positive aspects of the wrapper approaches consist in the consideration of interdependences and correlations among variables and in the fact that they focus directly on optimizing the performance of the learning algorithm considering also the bias of the prediction algorithm [7]. Due to all these reasons wrappers generally outperform filters [2] [38]. However the stability is a critical parameter for evaluating the performance of the wrappers, as they show a high sensitivity with respect to variations of the data that are used for the training procedure. The generic scheme of a wrapper is shown in figure 1.
Some of the more commonly used Wrapper strategies could be the following:

- **Exhaustive Search** The simplest example of wrapper is provided by the analysis of all combinations of variables, the so–called Exhaustive Search (ES) or *brute force* method. Such method takes into consideration all the possible combination of input variables, for each of them a classifier is trained and its accuracy is computed: the variable combination providing the highest classification accuracy is finally selected [34]. This approach is in principle the most reliable one, but its application is limited as its computational time complexity is exponential and it is also quite unstable.
- **Greedy search approach** The Greedy search approach consists in iteratively adding or removing features from the data to select a feature subset that maximizes the

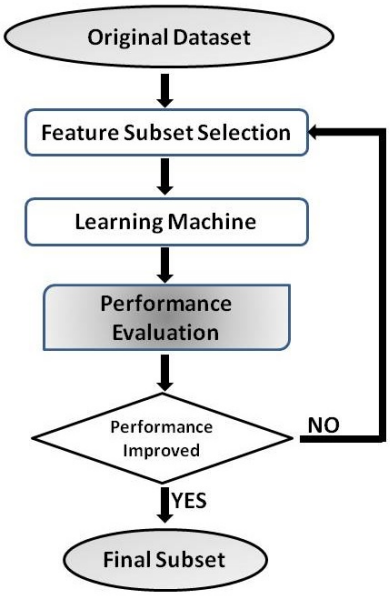

Figure. 1: General Wrapper scheme

accuracy of the learning algorithm. The most common search strategies are Sequential Forward Selection (SFS) and Sequential Backward Selection (SBS). SFS starts with an empty set of variables and iteratively increments the features set until a stopping criterion is satisfied. The search stops when the addition of new variables to the input set does not improve the performance of the model or classifier. SBS starts with an input set containing all the available features and eliminates them one by one. The relevance of an input variable is determined by removing one of them and calculating the performance of the classifier without having such variables among its inputs. The search stops when deleting features from the input set lead to a decrease of the performances. SFS is less expensive than SBS approach because it stops early [9]; SBS becomes impracticable when the number of potential input features is too large.

- **Metaheuristics algorithm.** This category includes local and global metaheuristic algorithms such as tabu search or evolutionary algorithms [39]. An example of wrapper based on evolutionary computation is provided in [7], where a Genetic Algorithm (GA)-based procedure (GAW) suitable to classifiers is proposed. The GA chromosomes are binary, their length corresponds to the number of input variables and each gene is associated to an input. A chromosome corresponds to the particular subset of variables whose corresponding gene has a unitary value. Initially a set of possible solutions (also named *population* in the GA terminology) is randomly generated and the so-called *fitness* of each of them is computed, which corresponds to the associated performance of the classifiers. The solutions showing the lowest fitness values are discarded, while the ones showing the highest fitness values are retained and exploited to generate new solutions through the crossover and mutation operations. The crossover operation creates new chromosomes (*sons*) by mating two existing ones (*parents*), e.g. by randomly selecting the genes values from the two parents. The mutation operation generates new solutions starting from one single solution, e.g. by randomly selecting a switching one or more genes of the original chromosome. The stop conditions for the algorithm often consist in the achievement of a fixed number of iterations or of a stable value of the average the fitness value of the whole population.

Embedded approaches are similar to wrappers, as they involve the learning algorithm in the variables selection procedure but, rather than repeatedly using the learning algorithm as a black box on each candidate subset, these procedures integrate a pre-selection step as part of the model training process. Embedded approaches need iterative updates and the evolution of the process parameters consider the efficiency of the model under process [40] [41] [42] [43] [44].
A schematic comparison among the three above described feature selection approaches including their essential characteristics is shown in Table 1 [45].

III. The stability problem in classifiers design

The stability concept was firstly introduced in 1995 by Turney [53] and consists in the sensitivity of a classifier, which is designed through a training procedure exploiting a dataset, with respect to variations in the training dataset. In effect several studies demonstrate that the exploitation of different training sets can lead the same variable selection procedure to select quite different variable subsets. Turney proposed a method based on the agreement of the classification methods obtained by the same algorithm trained on different datasets. The agreement of two classification methods is defined as the probability that they provide the same results over all possible sample input vectors drawn from a given probability distribution [54].
Other contributions on stability problem were based on the bias-variance decomposition of the error of the classification models [55] and [56]. The variance term measures instability of the classification algorithms, as it quantifies the percentage of times that the outputs provided by the classifiers trained with different training sets for a certain instance are different from the typical one. Bias-variance decomposition is normally performed via bootstrapping, where a portion of the data is used as test set while the remaining one is exploited to build different training sets through sampling with replacements. The final variance estimate corresponds to the average value which is computed over the different boot strapped samples.
Somol et al. introduced a method based on Shannon Entropy [57] for evaluating the similarity of variable subset selectors but they did not provide any method for establishing the stability. Moreover they did not develop tests on any real world data sets .
Kalousis et al. [17], in order to evaluate stability by assessing the difference in the results provided by two runs of the same variable selection procedures exploiting different training datasets, applied some indexes deriving from statistics, such as the Tanimoto distance, the Spearman rank correlation coefficient and the Pearson correlation coefficient. The obtained results demonstrated that none of considered variable

Table 1: Overview of the three main cathegories of feature selection techniques.

Model	Advantages	Disadvantages	Examples
Filter	Fast	No features dependencies	Chi-square [36]
	Scalable	No interaction with the learning algorithm	Euclidean distance [46]
	Independent of the classifier		t-test [47]
			Information Gain [37]
			Correlation based feature selection [35]
Wrapper	Simple	Risk of over fitting	Sequential forward selection [48]
	Interaction with the learning algorithm	More prone than randomize algorithms	Sequential backward elimination [49]
	Models feature dependencies	classifier dependent selection	Genetic algorithms [5]
Embedded	Interaction with the learning algorithm	Classifier dependent selection	Decision trees [50]
	Better computational complexity		Weighted naive Bayes [51]
	Models feature dependencies		feature Selection using the weight vector of
			Support Vector Machine (SVM) classifier [52]

selection methods are stable for the datasets which where analysed in his work. Moreover Kalousis observed that instability more frequently occurs when the initial global set of features shows a high level of redundancy [17] [58]. Some outcomes of this work will be depicted in the following subsection and exploited in the present work.

A. *Stability Measures*

Let us consider a classification problem where n potential input variables need to be considered for variable selection: the training samples can thus be described by a vector of n variables $\vec{v} = (v_1, v_2, ..., v_n)$. There are three kinds of representations in which variable selection approach can indicate feature preferences: in fact, the different variable selection methods usually provide their outcomes in one of the following forms:

- a weighting-scoring $\vec{w} = (w_1, w_2, ..., w_n)$, $w \in W \subseteq R^n$;
- a ranking vector: $\vec{r} = (r_1, r_2, ..., r_n)$, $1 \leq r_k \leq n$, where r_k represent the rank of variable k;
- an n-dimensional binary vector where each component is associated to a feature and its null or unitary value represents, respectively, absence or presence of a variable in the selected subset: $\vec{b} = (b_1, b_2, ..., b_n)$, $b_k \in [0, 1]$

In order to evaluate the stability of a variable selection method, a measure of similarity for each of the three representations must be preliminarly introduced. In the first case, in order to quantify similarity between two weighting vectors $\vec{w}^1 = (w_1^1, w_2^1, ..., w_n^1)$ and $\vec{w}_2 = (w_1^2, w_2^2, ..., w_n^2)$ the *Pearson's correlation coefficient* [59] can be calculated as follows:

$$S_w(\vec{w}^1, \vec{w}^2) = \frac{\sum_{k=1}^{n} (w_k^1 - \mu_1) \cdot (w_k^2 - \mu_2)}{\sqrt{\sum_{k=1}^{n} (w_k^1 - \mu_1)^2 \cdot \sum_k (w_k^2 - \mu_2)^2}} \quad (1)$$

where $\mu_1 = 1/n \sum_{k=1}^{n} w_k^1$ and $\mu_2 = 1/n \sum_{k=1}^{n} w_k^2$. S_w lies in the range $[-1, 1]$: the null value represents absence of correlation while unitary values mean that $\vec{w}_1$ and $\vec{w}_2$ are exactly (positively or negatively) correlated.
Similarly, in order to quantify the similarity between two rankings $\vec{r}^1 = (r_1^1, r_2^1, ..., r_n^1)$ and $\vec{r}^2 = (r_1^2, r_2^2, ..., r_n^2)$, the *Spearman's rank correlation coefficient* [60], which is also roughly indicated as "the Pearson correlation coefficient between the ranked variables", is computed as follows:

$$S_R(\vec{r}^1, \vec{r}^2) = 1 - 6 \cdot \frac{\sum_{k=1}^{n} (\vec{r}_k^1 - \vec{r}_k^2)}{m(m^2 - 1)} \quad (2)$$

S_R lies in the range $[-1, 1]$: the unitary value indicates that the two rankings are identical, the null value represents no correlation between the two ranks and the value -1 indicates that the two rankings have inverse orders.
Finally the similarity between two binary vectors $\vec{b}^1$ and $\vec{b}^2$ is evaluated through the *Tanimoto distance* [61], which is computed as:

$$S_B(\vec{b}^1, \vec{b}^2) = \frac{|\vec{b}^1 \cdot \vec{b}^2|}{|\vec{b}^1| + |\vec{b}^2| - |\vec{b}^1 \cdot \vec{b}^2|} \quad (3)$$

where $|\cdot|$ indicates the norm of the binary vector and $\vec{b}^1 \cdot \vec{b}^2$ is the scalar product of $\vec{b}^1$ and $\vec{b}^2$. S_B lies in the range $[0, 1]$ where the null value means that there is no overlap between the two sets while an unitary value represents that the two sets are identical [17].

IV. The proposed approach

The basic idea behind the approach proposed here lies in the consideration of the mutual interaction between couples of variables, which is usually neglected by standard variable selection procedures, considering that stability is mostly compromised when the available variables show a high level of redundancy [17] and this redundancy must be eliminated in order to gain a stable outcome.
Let us consider a dataset containing n potential input variables for an AI-based classifier. In the following the depicted strategy is applied to three different kinds of classifiers: Bayesian, Decision Tree (DT) and Linear Discriminant Analysis (LDA), in order to demonstrate that the proposed approach is generic and can be applied to any kind of binary classifier tuned through a supervised learning approach, which exploits a dataset for its design and test.
For a fixed number of times M the available dataset is shuffled and partitioned into 3 subsets: a Training Set (containing 60% of the available data), a Validation and a Test Set (each of these latter ones holding 20% of the data). The variable selection algorithm is run and the selected variables subset is recorded. Since variable selection procedures are deterministic, using the same training data would lead to the selection

of the same variable subset. Therefore the instability is generated by the fact that the samples in the training set vary at each run due to the preliminary shuffling step. Afterwards a new subset is built by including the couples of variables which are more frequently jointly selected by the variable selection procedure.
The reason for analysing couples and not bigger subsets, such as triples or quadruples, lies in the following reasoning: let us suppose that a triple of variables (v_1, v_2, v_3) is selected p times. Then, due to the so-called *a-priori property* [62], any of the 3 possible subsets (i.e. couple) of this triple (v_1, v_2), (v_1, v_3), (v_2, v_3) needs be selected at least p times. If one of the three couples is selected z times (with $z > k$), it is preferable to consider that couple instead of the whole tripe, as its information content with respect to the classification problem shows to be greater. Thus looking at the couples avoids to omit any larger frequent subsets. On the other hand, the analysis of the couples of variables rather than of each single feature (e.g. through a simple histogram), allows to assess the interaction between the variables.
The proposed approach can be summarised as follows:

1. For M times the available dataset is shuffled and partitioned into 3 subsets: Training Set (containing 60% of the available data), Validation and Test Sets (each of these latter ones holding 20% of the data),

2. the variable selection procedure is run on each of the M training datasets obtained in step 1): the accuracy of the trained classifier is calculated on the corresponding validation set in order to assess the goodness of the variable selector. The variables subsets $\vec{b}_i, (1 \leq i \leq M)$ which is identified by each run of the variable selection procedure is stored as a row of a binary $M \times n$ matrix $\mathbf{B}$.

3. Using the matrix $\mathbf{B}$ the couples of features which are most frequently jointly selected by the variable selection approach are identified. Among all the possible $n(n-1)/2$ couples of variables, the ones which have been selected for the highest number of times are identified.

4. A new variables subset $\vec{q}$ is built by merging the previously identified couples of variables.

5. The M triples of Training, Validation and Test datasets that have been built at step 1) are reduced by considering only the variables selected at step 4)

6. The M reduced training datasets built at step 5) are fed as input to the selected classifier and the corresponding test sets are used to evaluate the classifier accuracy.

   The accuracy is calculated using the so called Balanced Classification Rate (BCR) [63] that considers the balance between the two class in order to be an appropriate measure for balanced and imbalanced data [64]. [65]. BCR is computed as follows:

$$BCR = \frac{1}{2} \cdot \left[\frac{TP}{TP + FN} + \frac{TN}{TN + FP} \right] \quad (4)$$

   where *TP (True Positive)* represents the percentage of correctly classified unitary samples; *TN (True Negative)* represents the percentage of correctly classified null samples; *FP (False Positive)* represents the percentage of null samples incorrectly classified and finally *FN (False Negative)* represents the percentage of unitary samples incorrectly classified.

7. The classifier showing the highest BCR is finally selected.

The procedure is schematically depicted in Figure 2.

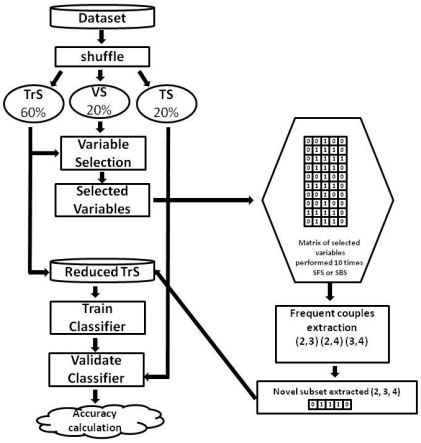

Figure. 2: General scheme

V. Experimental tests

The proposed variable selection approach is generic and can be performed on any kind of binary classifier tuned through a supervised learning approach exploiting a dataset for its design. In the following the depicted strategy has been applied to build three different classifiers (Bayesian, DT-based and LDA-based) applied to different datasets extracted by the largely used UCI learning repository [66] as well as to two datasets coming from industrial field. A deep comparison with the standard SFS, SBS, GAW or ES procedures has been developed. The exploited datasets are described below and a summary of their main features are summarized in table 2.

- **Breast Cancer Wisconsis (BCW)**: the BCW database is provided by the Hospital of the University of Wisconsis. Data refer to about 700 patients affected by tumors. The binary target describes if the tumor is benign or malign.

- **Australian Credit Approval (ACA)**: data contains details about credit card applications including continuous and nominal attributes. The binary target indicates if the customer has a good or bad credit.

- **Mammography Mass (MM)**: Mammography mass dataset is collected by the Institute of Radiology of the Erlangen-Nuremberg University during 2003-2006. The binary target is used to classify the mass lesion as benign or malign.

- **Pima Indians Diabetes (PID)**: this dataset is an extraction of a bigger dataset which is held by the National Institutes of Diabetes and Digestive and Kidney Diseases. This analysis was carried out on women patients at least 21 years old coming from Arizona. The target is used to establish if the diabetes test is positive or negative.

- **Heart**: This dataset is built to classify the presence or absence of heart disease. This database is extracted by another larger dataset containing 76 attributes, but all published experiments refer to using a subset of 14 of them.

- **Monk2**: The Monk's problems consist of a set of three artificial problems including the same set of features. The three problems come from an artificial robot and differ on the kind of concept to be learned. Moreover the differences include also the amount of noise in the training set. Each problem is provided by a logical characterization of a concept and robots can or cannot appertain to this concept.

- **Blood Transfusion Service Center (BTSC)**: Blood Transfusion Service Center Data Set used the donor database of Blood Transfusion Service Center in Hsin-Chu City in Taiwan. The dataset is built by extracting 748 donors at random from the donor database. The target indicates whether a donor donated blood in March 2007.

- **Industrial I**: This dataset comes from an industrial context: it represents the outcome of the quality control and contains products analyses. The binary target indicate if a product is defective and must be discarded or non-defective and can be send to the market.

- **Industrial II**: This dataset belong to the metal industry field. The input variables depict operational parameters and it is a good example of real imbalanced dataset. The minority class represents the occurrence of the faulty situation and to detect these rare cases is very important.

Table 2: Dataset description

Dataset	#Instances	#Features	#Class0	#Class1
BCW	699	9	458	241
ACA	690	14	383	307
MM	830	5	427	403
PID	768	8	500	268
Heart	270	13	120	150
Monk2	432	6	204	228
BTSC	748	4	570	178
Industrial I	1235	26	517	718
Industrial II	1915	10	1454	461

In order to assess whether the depicted procedure improves the stability of the variable selection procedures, such procedures are repeated Q times, by thus identifying Q variables subsets, each represented through a binary vector $\vec{p}_i$ ($1 \leq i \leq Q$). The stability is quantified by the average Tanimoto distance $\bar{T}$ among the binary vectors $\vec{p}_i$, that can be directly compared to the average Tanimoto distance $\bar{T}_{ES}$, $\bar{T}_{SFS}$, $\bar{T}_{SBS}$ or $\bar{T}_{GAW}$ among the binary vectors $\vec{b}_i$ obtained in Q runs of the standard wrapper procedures.

In order to compare the performance of the designed classifiers, the average $B\bar{C}R$ of the Q accuracy values obtained for the classifier on each test set can be used as overall performance index for the classifier that is built by means of proposed approach and compared to the average $B\bar{C}R_{ES}$, $B\bar{C}R_{SFS}$, $B\bar{C}R_{SBS}$ or $G\bar{A}W_{SBS}$ of the Q accuracy values shown by classifier using the M variable subsets selected by the standard ES procedures, SFS, SBS and GAW, respectively. Finally the number of selected variables has also been considered a meaningful term of comparison among standard and the proposed approach: thus the average number $\bar{N}_{ES}$, $\bar{N}_{SFS}$, $\bar{N}_{SBS}$ or $\bar{N}_{GAW}$ and $\bar{N}$ of selected variables in the Q trials of the wrappers and the proposed approach, respectively, have also been computed.
Within the tests discussed here, for the proposed procedure $M = 10$ have been set. Moreover the comparison among the two approaches have been developed on $Q = 10$ independent runs.

A. Results for the Exhaustive Search

Table 3 shows the results obtained in the tests developed on all the proposed datasets by applying the ES procedure. In particular the first column indicates the processed dataset, columns 2-4 represent the values of the selected indexes for the standard ES procedure, while the columns 5-7 represent the values of the selected indexes for the proposed approach.

Table 3: Results of the tests with ES

Dataset	$B\bar{C}R_{ES}$	$\bar{N}_{ES}$	$\bar{T}_{ES}$	$B\bar{C}R$	$\bar{N}$	$\bar{T}$
Bayesian Classifier						
BCW	0.95	4.9	0.4	0.96	2.7	0.6
ACA	0.86	2.5	0.45	0.87	2.4	0.6
MM	0.76	2.8	0.47	0.78	2.5	0.63
PID	0.70	4.9	0.52	0.71	2.4	0.60
Heart	0.76	6.9	0.40	0.76	2.7	0.44
Monk2	0.94	3.6	0.6	0.96	2	1
BTSC	0.66	2.7	0.6	0.68	2.1	0.66
Industrial I	0.72	5.9	0.5	0.72	2	1
Industrial II	0.73	3.9	0.5	0.83	2.3	0.65
LDA-based Classifier						
BCW	0.95	5.7	0.55	0.95	2.6	0.61
ACA	0.86	7.4	0.47	0.88	2.8	0.60
MM	0.80	3.8	0.69	0.80	2.6	0.74
PID	0.73	5.8	0.68	0.75	2.9	0.71
Heart	0.81	9.4	0.61	0.83	2.6	0.68
Monk2	0.78	3.4	0.58	0.80	2	1
BTSC	0.60	2.7	0.68	0.63	2.6	0.72
Industrial I	0.80	5.5	0.63	0.81	3	1
Industrial II	0.73	3.3	0.60	0.77	2.3	0.70
DT-based Classifier						
BCW	0.93	4.4	0.52	0.94	3.3	0.69
ACA	0.82	7.1	0.4	0.88	3.1	0.5
MM	0.81	2.9	0.52	0.83	2.3	0.56
PID	0.66	4.9	0.48	0.78	2.1	0.51
Heart	0.76	6.6	0.39	0.78	2.1	0.51
Monk2	0.78	3.4	0.55	0.79	2	1
BTSC	0.59	2.8	0.58	0.66	2.3	0.61
Industrial I	0.79	6.4	0.58	0.79	2.5	0.82
Industrial II	0.71	3.7	0.5	0.74	2.6	0.60

Considering the obtained results, as far as the ES is concerned, it can be concluded that:

- In many cases the proposed approach improves accuracy and stability by also reducing the number of selected input variables. However there are some examples

where the proposed method does not improve the mean accuracy of the classifier but reduces the number of selected variables and increases the stability. In some examples, such as **Monk2** and **Industrial I** the stability shows even an unitary value.

- On average, the best mean accuracy of the proposed approach, in terms of absolute value is obtained with the DA-based classifier, but, if the percentage improvement with respect to the the classical ES is considered, then the best improvement has been obtained with the DT-based classifier (6.1%).
- As far as the mean subset length and the stability are concerned, the best results are obtained with the LDA-based classifier considering both the absolute values and the percentage improvement with respect to the traditional method, about 33% and 26% respectively.

B. Results for the Sequential Forward Selection

Table 4 shows the results obtained in the tests developed on all the proposed datasets by applying the SFS procedure for the three selected classifiers.

Table 4: Results of tests with SFS

Dataset	$B\bar{C}R_{SFS}$	$\bar{N}_{SFS}$	$\bar{T}_{SFS}$	$B\bar{C}R$	$\bar{N}$	$\bar{T}$
Bayesian Classifier						
BCW	0.94	5.1	0.6	0.94	2.1	0.93
ACA	0.85	2.4	0.45	0.87	2.4	0.6
MM	0.78	2.3	0.56	0.80	2.3	0.77
PID	0.71	3.7	0.59	0.71	2.4	0.65
Heart	0.75	4.6	0.39	0.8	2.9	0.84
Monk2	0.95	2.4	0.8	0.96	2	1
BTSC	0.64	2.7	0.65	0.66	2.1	0.73
Industrial I	0.82	3.2	0.53	0.83	2.5	0.7
Industrial II	0.7	3.4	0.7	0.8	2.1	0.9
LDA-based Classifier						
BCW	0.93	4.5	0.52	0.95	2.3	0.70
ACA	0.86	2.4	0.77	0.87	2	1
MM	0.80	1.4	0.52	0.82	1.4	0.77
PID	0.73	4.4	0.52	0.73	2.3	0.76
Heart	0.78	4.9	0.39	0.82	2.5	0.64
Monk2	0.78	2.2	0.58	0.80	2	1
BTSC	0.60	2.7	0.68	0.62	2.6	0.72
Industrial I	0.80	4.9	0.65	0.81	3	1
Industrial II	0.70	4.4	0.70	0.72	2.9	0.72
DT-based Classifier						
BCW	0.94	3.1	0.4	0.96	2.1	0.73
ACA	0.85	2.4	0.51	0.86	2.4	0.59
MM	0.82	2.2	0.55	0.85	2.2	0.76
PID	0.66	3.3	0.37	0.78	2.7	0.47
Heart	0.75	2.9	0.40	0.81	2.6	0.62
Monk2	0.96	2.9	0.82	0.99	2.6	0.95
BTSC	0.59	3	0.73	0.68	2.4	0.73
Industrial I	0.79	5.1	0.65	0.81	3	1
Industrial II	0.68	4.1	0.64	0.73	3.1	0.74

Considering the obtained results, as far as the SFS is concerned, it can be concluded that:

- The proposed approach, despite in some cases having the same accuracy, improves the stability until 55% , by also reducing the number of selected input variables. This represents a good result, as the novel method extracts less variables which are actually those ones which mainly affect the considered target, as shown by the increased stability.
- There are some examples where the mean length of the selected subsets is the same for the two approaches but the proposed method improves both the average accuracy and the mean stability.
- In many cases the proposed approach leads to an improvement of all the considered performance indexes for all the database.
- On average, the best mean accuracy of the proposed approach, in terms of absolute value is obtained by the DT-based classifier, but, if the percentage improvement with respect to the the classical SFS is considered, then the best improvement has been obtained by the LDA-based classifier (13%).
- As far as the mean subset length and the stability are concerned, the best results are obtained by the LDA-based classifier considering both the absolute values and the percentage improvement with respect to the traditional method.

C. Results for the Sequential Backward Selection

Table 5 shows the results obtained in the tests related to the SBS procedure.

Table 5: Results of the tests with SBS

Dataset	$B\bar{C}R_{SBS}$	$\bar{N}_{SBS}$	$\bar{T}_{SBS}$	$B\bar{C}R$	$\bar{N}$	$\bar{T}$
Bayesian Classifier						
BCW	0.95	7.1	0.48	0.98	2.8	0.67
ACA	0.85	7.7	0.59	0.86	2.3	0.76
MM	0.79	3.5	0.65	0.79	2.4	0.78
PID	0.72	5.1	0.65	0.73	2.6	0.73
Heart	0.77	9.3	0.43	0.79	3.2	0.64
Monk2	0.94	2.8	0.70	0.95	2	1
BTSC	0.65	3.2	0.75	0.67	2.2	0.89
Industrial I	0.71	5.5	0.52	0.73	2	1
Industrial II	0.75	3.3	0.68	0.78	2.6	0.83
LDA-based Classifier						
BCW	0.93	5.7	0.55	0.95	2.6	0.61
ACA	0.86	7.4	0.47	0.87	2.8	0.60
MM	0.80	3.8	0.69	0.80	2.6	0.74
PID	0.73	5.8	0.68	0.73	2.9	0.71
Heart	0.81	9.4	0.61	0.81	2.6	0.68
Monk2	0.78	3.4	0.58	0.81	2	1
BTSC	0.60	2.7	0.68	0.62	2.6	0.72
Industrial I	0.80	5.5	0.63	0.80	3	1
Industrial II	0.73	3.3	0.60	0.77	2.3	0.70
DT-based Classifier						
BCW	0.93	4.4	0.4	0.95	2.7	0.4
ACA	0.82	7.8	0.45	0.87	2.5	0.57
MM	0.81	2.5	0.58	0.84	2.1	0.93
PID	0.67	5.9	0.62	0.79	2.5	0.62
Heart	0.74	6.5	0.42	0.84	3.3	0.52
Monk2	0.99	2.9	0.82	0.99	2.6	0.95
BTSC	0.59	2.6	0.62	0.67	2.3	0.80
Industrial I	0.80	6.9	0.62	0.83	4.6	0.70
Industrial II	0.70	2.4	0.71	0.73	2.2	0.73

The results obtained for the SBS allow to conclude that:

- Sometimes the proposed method is not able to improve the mean accuracy of the classifier but reduces the number of selected variables and improves the stability.

However in some example the stability reaches even a unitary value, i.e the same variables are selected when changing the training set. Finally in most cases the proposed algorithm improves all considered indexes with respect to the traditional SBS algorithm.

- On average, the best mean accuracy, in terms of both absolute value and the percentage improvement (with respect to the the classical SBS) is obtained by the proposed approach by the DT-based classifier;
- the smallest mean selected subset length and the best average stability are obtained by the Bayesian classifier.

D. Results for the GAW approach

Table 6 shows the results obtained in the tests developed on the GAW procedure.
The results obtained for the GAW method show that:

- Again, there are cases where the accuracy does not improve even if the other indexes improve and in most of cases the proposed approach lead to an improvement of all the considered performance indexes.
- On average, the best mean accuracy , in terms of absolute value is obtained by applying the proposed procedure to the DT-based classifier: the percentage improvement which is about 8%.
- As far as the mean subset length is concerned, the best results, in terms of percentage improvement (47%) with respect to the standard approach, is obtained by the LDA-based classifier.
- The best results in terms of stability are obtained by the proposed approach when applied to the bayesian classifier.

E. General overview of the results

In all considered examples, the proposed approach improves the stability of all the considered variable selection procedures applied to different kind of classifier and processing different datasets. The novel procedure is indeed more time consuming with respect to the standard wrapper approaches, as each of its run requires the construction of M datasets, M runs of the algorithm and other accessory steps. However, the variable selection procedure is very often run once or at least at a low frequency and in most of the cases off-line, therefore this increased computational effort is in most cases sustainable.
Table 7 shows the average percentage improvements of the three considered indexes (computed on all the datasets and indicated as $\bar{\Delta}_{BCR}$, $\bar{\Delta}_N$ and $\bar{\Delta}_T$, respectively) that have been achieved with respect to the traditional approach by adopting the proposed method for each tested variable selection method.
All considered indexes, in average, are improved by applying the proposed approach. In particular, although the mean uncertainty in the classification slightly rises, the length of the selected subset of variables shows a satisfactory reduction and also the stability is improved. This means that, even when the performance of the classifier remains unchanged, the proposed method guarantees a reduced subset of variables and a greater stability of the selection of such variables. The highest percentage increase in terms of classifier accuracy is obtained when applying the GAW variable selection approach and the DT-based classifier. Moreover, as far and the number of selected variables is concerned, the smallest average value is achieved for the GAW variable selection method and the LDA-based classifier. Finally, the highest average percentage increase in stability is achieved for GAW and the Bayesian classifier. What above shows that the proposed approach is powerful when applied to a meta-heuristic wrapper variable selection approach, probably due to the highest sensitivity of these kind of method with respect to the selection of the training dataset. It can be noted, however, that the there are not important variations between the different variable selection methods, which also confirms that the efficiency of the proposed procedure is independent on the variable selection method used.
In order to analyse the effect of the proposed approach with respect to the type of the binary classifier, the average percentage improvements of the considered performance indexes over the different variable selection procedures (indicated in the following as $\bar{\delta}_{BCR}$, $\bar{\delta}_N$ and $\bar{\delta}_T$, respectively) have been computed for each classifier and each dataset analysed. Table 8 depicts the results.
It can be observed that the average percentage increase in terms of classifiers accuracy is quite low in most cases; however the highest improvements are obtained by the DT-based classifier. Concerning the length of the selected variables subset, for many datasets a considerable reduction is achieved up to 60%. Such reduction is clearly higher for dataset including many input variables, as the probability of having redundant or useless variables is higher. Finally it is worth noting that the value of the stability index does not depend neither on the classifier nor on the dataset and the proposed method for all the classifiers leads to an improvement of the stability.

Table 7: Percentage improvement of computed indexes for the different variable selection methods and classifiers

Index	classifier	**ES**	**SFS**	**SBS**	**GAW**
$\bar{\Delta}_{BCR}$	Bayesian	1.27	7.2	2.1	3.6
	LDA-based	0.46	3.5	1.6	7.7
	DT-based	6.1	4.7	6.5	7.9
$\bar{\Delta}_N$	Bayesian	33.4	25.5	46.8	40.1
	LDA-based	44.8	27.5	43.1	47
	DT-based	12.3	19.3	32.6	17.8
$\bar{\Delta}_T$	Bayesian	25.3	23.6	24.7	28.7
	LDA-based	25.6	26.4	16.7	25.8
	DT-based	21.6	22.1	14.2	23.5

VI. Conclusion and future work

The paper proposes a novel approach that increases the stability of some wrapper variable selection algorithms when applied to binary classification tasks. In fact the main problem of these algorithms lies in their high sensitivity to the variation of the training set. This represents a serious problem, especially when the main aim of the variable selection is knowledge extraction on the considered phenomenon, namely the selection of the variables that mainly affect the classi-

Table 6: Results of the tests with GAW

Dataset	$B\bar{C}R_{GAW}$	$\bar{N}_{GAW}$	$\bar{T}_{GAW}$	$B\bar{C}R$	$\bar{N}$	$\bar{T}$
Bayesian Classifier						
BCW	0.93	4.7	0.4	0.93	2.5	0.7
ACA	0.85	2.5	0.5	0.90	2.2	0.7
MM	0.77	2.8	0.5	0.78	2.3	0.8
PID	0.69	4.9	0.5	0.70	2.5	0.6
Heart	0.75	6.9	0.4	0.76	2.8	0.5
Monk2	0.94	3.6	0.5	0.97	2	1
BTSC	0.67	2.7	0.7	0.69	2.1	0.8
Industrial I	0.73	5.9	0.6	0.74	2.2	0.9
Industrial II	0.72	3.9	0.5	0.85	2.4	0.6
LDA-based Classifier						
BCW	0.94	5.1	0.43	0.94	2.3	0.53
ACA	0.86	7.7	0.42	0.86	2.4	0.45
MM	0.78	3.2	0.5	0.80	2.4	0.7
PID	0.72	4.9	0.5	0.72	2.3	0.6
Heart	0.81	7.5	0.5	0.81	2.7	0.7
Monk2	0.77	3.8	0.5	0.79	2	1
BTSC	0.7	2.6	0.7	0.8	2.1	0.8
Industrial I	0.80	6.7	0.6	0.80	2.8	0.8
Industrial II	0.73	3.7	0.5	0.75	2.5	0.7
DT-based Classifier						
BCW	0.94	3.1	0.4	0.96	2.1	0.73
ACA	0.85	2.4	0.51	0.86	2.4	0.62
MM	0.82	2.2	0.55	0.83	2.2	0.76
PID	0.6	3.3	0.4	0.8	2.7	0.5
Heart	.0.8	2.9	0.40	0.8	2.6	0.6
Monk2	0.96	2.9	0.8	0.99	2.6	0.94
BTSC	0.6	3	0.74	0.7	2.4	0.8
Industrial I	0.78	5.1	0.65	0.82	2.8	0.95
Industrial II	0.70	4.1	0.64	0.74	3.1	0.74

Table 8: Mean percentage improvement of the proposed method for each dataset and each classifier

	$\bar{\delta}_{BCR}$			$\bar{\delta}_N$			$\bar{\delta}_T$		
Dataset	**Bayesian**	**LDA-based**	**DT-based**	**Bayesian**	**LDA-based**	**DT-based**	**Bayesian**	**LDA-based**	**DT-based**
BCW	1	1.1	2.1	52.8	49.9	33.8	34.8	17.6	33.8
ACA	2.5	0.56	2.3	31.8	53.9	33.7	25.3	18.5	16.5
MM	1.6	1.2	21.8	24	18.5	4	25.1	21.1	28.5
PID	1.1	0	17.8	45.7	50.9	28.1	12.5	21.2	25
Heart	1.2	1.5	6.7	55.7	59.5	20	26.7	27.8	30.9
Monk2	3.1	1.6	2.3	25.2	36.2	8.5	31.1	45	11.4
BTSC	2.95	1.7	12.7	24.5	12.9	17.9	12.1	9	7.5
Industrial I	1.3	0.3	5.7	53.5	49.3	30.5	38.9	41	19.5
Industrial II	7.9	10.5	5.7	34.7	32.4	18.7	20.7	15.6	10.8

fication. A stable variable selection algorithm allows a more efficient identification of the factors which are most relevant with respect to the classification problem. The proposed method is applied to the design of three different classifiers: bayesian classifier, the LDA-based classifier and DT-based classifier. Moreover several datasets coming from real word applications have been processed. The obtained results show that the proposed approach is effective independently on the type of variable selection method, on the type of classifier and on the database.
Future work will deal with the extension of this method in order to make it suitable to cope with tasks which are different from classification, such as development of AI-based approximation or forecasting models or clustering.

References

[1] I. Guyon, A. Elisseeff. An introduction to variable and feature selection, *Machine Learning*,3, pp. 1157-1182, 2003.

[2] R. Kohavi, G. John. Wrappers for feature selection, *Artificial Intelligence*, 97, pp. 273-324, 1997.

[3] H. Liu, H. Motoba, R.Setiono, Z. Zhao. Feature selection: an ever evolving frontier in data mining. In *JMLR: workshop and Conference proceedings: the 4th workshop on Feature Selection in Data Mining*, pp. 4-13, 2010.

[4] M. Dash, H.Liu. Feature selection for classification, *Intelligent Data Analysis*,1, pp. 131-156, 1997.

[5] S. Cateni, V. Colla, M. Vannucci. General purpose input variable extraction: a genetic algorithm based procedure give a gap, in *9th International Conference on Intelligence Systems design and Applications ISDA'09*, pp. 1307-1311, 2009.

[6] D. Sofge, D.Elliot. Improved neural modelling of real world systems using gengene algorithm based variable selection. In *proc. Conference on Neural Networks and Brain*, pp. 1-4, 1998.

[7] S. Cateni, V. Colla, M. Vannucci. Variable selection through genetic algorithms for classification purpose. In *IASTED International Conference on Artificial Intelligence and Applications AIA2010*, pp. 6-11, 2010.

[8] S. A. Lashari, R. Ibrahim, N. Senan. Fuzzy soft set based classification for mammogram images. *International Journal of Computer Information Systems and Industrial Management Applications*, 7(1), pp. 66-73, 2015.

[9] S. Cateni, V. Colla, M. Vannucci. A genetic algorithm based approach for selecting input variables and setting relevant network parameters of SOM based classifier, *International Journal of Simulation Systems Science and Technology*,12 (2), pp. 30-37, 2011.

[10] S. Wang, J. Zhu. Variable selection for model-based high dimensional clustering and its application on microarray data. *Biometrics*,64, pp. 440-448, 2008.

[11] L.I. Kuncheva. A stability index for feature selection. In *IASTED International conference on Artificial intelligence and Application AIA2007*, pp. 95-116, 2007.

[12] A. Jain, D.Zongker. Feature selection: evaluation, application and small sample performance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*,19, pp. 153-158, 1997.

[13] C. Duan, Z. Fei, J. Li. A variable selection aided residual generator design approach for process control and monitoring. *Neurocomputing*, 171, pp. 1013-1020, 2016.

[14] V. Kovalishyn, G. Pod. Efficient variable selection batch pruning algorithm for artificial neural networks. *Chemometrics and Intelligent Laboratory Systems*, 149, pp. 10-16, 2015.

[15] X. Wang, M. Wang. Variable selection for high-dimensional generalized linear models with the weighted elastic-net procedure. *Journal of Applied Statistics*, 43 (5), pp. 796-809, 2016.

[16] A. Jain, B. Chandrasekaran. Dimensionality and sample size considerations in pattern recognition practice. *Handbook of statistics* ,1, pp. 835-855, 1982.

[17] A. Kalousis, J. Prados, M. Hilario. Stability of feature selection algorithms. In *Proc. 5th IEEE International Conference on Data Mining (ICDM'05)*, pp. 218-225, 2005.

[18] H. Eid, A. Hassanien, T. H. Kim, S. Banerjee. Linear correlation-based feature selection for network intrusion detection model, *Communications in Computer and Information Science*, 381 , pp. 240-248, 2013.

[19] L. Koc, A. D. Carswell. Network intrusion detection using a HNB binary classifier. In *17th UKSIM-AMSS International Conference on Modelling and Simulation*, pp.81-85, 2015.

[20] M. Shetty, N.M.Shekokar. Data mining techniques for real time intrusion detection systems, *International Journal of Scientific & Engineering Research*,3 (4), pp. 1-7, 2012.

[21] S. Ghumbre, C. Patil, A. Ghatol. Heart disease diagnosis using support vector machine. In *International Conference on Computer Science and Information Technology (ICCSIT'2011) Pattaya*, pp.84-88, 2011.

[22] M. Khashei, S. Eftekhari, J. Parvizian. Diagnosing diabetes type II using a soft intelligent binary classification model, *Review of Bioinformatics and Biometrics (RBB)*,1, pp. 9-23, 2012.

[23] W. Froelich, K. Wrobel, P. Porwik. Diagnosing parkinson's disease using the classification of speech signals. *Journal of medical informatics and technologies*, 23, pp. 187-194, 2014.

[24] L. Yu. Redundancy based feature selection for microarray data. In *Proc. of the tenth ACM SIGKDD Conference on Knowledge Discovery and Data mining*, 1, 2004.

[25] X. Zhao, L. W. Cheung. Kernel-embedded gaussian processes for disease classification using microarray gene expression data, *BMC Bioinformatics*, 8 (67), pp. 1-27, 2007.

[26] X. Jiang, B. Cai, D. Xue, X. Lu, G. F. Cooper, R. E. Neapolitan. A comparative analysis of methods for predicting clinical outcomes using high-dimensional genomic datasets, *Journal of the American Medical Informatics Association*,21, pp. 312-319, 2014.

[27] A. Nikooienejad, W. Wang, V. E. Johnson. Bayesian variable selection for binary outcomes in high dimensional genomic studies using non-local priors, *Bioinformatics*, 32 (2), 2016.

[28] S. Cateni, V. Colla. Improving the stability of sequential forward and backward variables selection. In *15 th International Conference on Intelligent Systems design and applications ISDA 2015, Marrakesh, Morocco.*, 2015.

[29] M. Sebban, R. Nock. A hybrid filter/wrapper approach of feature selection using information theory, *Pattern Recognition*,35, pp. 835-846, 2002.

[30] S. Zhang, Z. Zhao. Feature selection filtering methods for emotion recognition in chinese speech signal. In *9th International Conference on Signal Processing, ICSP*, pp. 1699 - 1702, 2008.

[31] P. L. Carmona, J. M. Sotoca, F. Pla. Filter-type variable selection based on information measures for regression tasks, *Entropy*, 14, pp. 323-343, 2012.

[32] L. Loo, S. Roberts, L. Hrebien, M. Kam. New filter-based feature criteria for identifying differentially expressed genes. In *Proc. of the Fourth International conference on Machine Learning and Applications*, pp.1-10, 2005.

[33] W. Zhao, R. Zhang. Variable selection of varying dispersion student-t regression models. *Journal of Systems Science and Complexity*, 28, pp. 961-977, 2015.

[34] S. Cateni, V. Colla, M.Vannucci. A hybrid feature selection method for classification purposes. In *8th European Modeling Symposium on Mathematical Modeling and Computer simulation EMS2014*, 1, pp. 1-8, 2014.

[35] L. Yu, H. Liu. Feature selection for high dimensional data: a fast correlation based filter solution. In *Proc. of the 20th International Conference on Machine Learning ICML*, pp. 856-863, 2003.

[36] S. Wu, P. Flach. Feature selection with labelled and unlabelled data. In *Proceeding of ECML7PKDD'02 Workshop on Integration and Collaboration Aspects of Data Mining Decision Support and Meta-Learning*,1, pp. 156-167, 2002.

[37] D. Roobaert, G. Karakoulas, V. Nitesh, V. Chawla. Information gain, correlation and support vector machines, *Studies in Fuzziness and Soft Computing*, 207, pp. 463-470, 2006.

[38] T. Golub, D. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. Mesirov, H. Coller, M. Loh, J. Downing, M. Caligiuri. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*,286, pp. 531-537, 1999.

[39] P. Bermejo, J. Gamez, and J. Puerta. A grasp algorithm for fast hybrid (filter-wrapper) feature subset selection in high dimensional datasets. *Journal Pattern Recognition Letters*, 32, pp. 701-711, 2011.

[40] Z. Yan, Y. Yao. Variable selection method for fault isolation using least absolute shrinkage and selection operator (LASSO). *Chemometrics and Intelligent Laboratory Systems*, 146, pp. 136-146, 2015.

[41] S. Kwon, S. Lee, Y. Kim. Moderately clipped LASSO. *Computational Statistics and Data Analysis*, 92, pp. 53-67, 2015.

[42] X. He, D. Cai, P. Niyogi. Laplacian score for feature selection, *Advances in Neural Information Processing Systems*, pp. 507-514, 2005.

[43] L. Bo, L. Wang, L. Jiao. Multilayer perceptrons with embedded feature selection with application in cancer classification. *Chinese Journal of Electronics*, 15, pp. 832-835, 2006.

[44] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, 58, pp. 267-288, 1996.

[45] Y. Saeys, I. Inza, P. Larranaga. A review of feature selection techniques in bioinformatics, *Gene expression Bioinformatics*, 23, pp. 2507-2517, 2007.

[46] E. Fowlkes, R. Gnanadesikan, J.Kettenring. Variable selection in clustering, *Journal of classification*, 5 (2), pp. 205-228, 1988.

[47] T. J. Mitchell J. J. Beauchamp. Bayesian variable selection in linear regression, *Journal of the American Statistical Association*, 83, pp. 1023-1032, 1988.

[48] D. Ververidis, C. Kotropoulos. Sequential forward feature selection with low computational cost. In *Proc. 13th European Signal Processing Conference, EUSIPCO 2005*, pp. 1063-1066, 2005.

[49] K. Z. Mao. Orthogonal forward selection and backward elimination algorithms for feature subset selection, *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics.*,34, pp. 629-634., 2004.

[50] P. E. Utgoff. Incremental induction of decision trees, *Machine learning*, 4, pp. 161-186, 1989.

[51] L. Breiman. Random forests, *Machine learning*, vol. 45, pp. 5-32, 2001.

[52] J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio, V. Vapnik. Feature selection for svms, *Advances in Neural Information Processing Systems*, 12, pp. 668-674, 2001.

[53] P. Turney. Techncal note:bias and the quantification of stability, *Machine Learning*, 20, pp. 23-33, 1995.

[54] Y. Han. Stable feature selection : theory and algorithms, *State University of New York at Binghamton, ProQuest Dissertations Publishing.*, 1, pp. 1-102, 2012.

[55] P. Domingosl. A unified bias-variance decomposition and its applications. In *Proc. 17th International Conference on Machine Learning*, 1, pp. 231-238, 2000.

[56] S. Geman, E. Bienenstock, R. Doursat. Neural networks and the bias/variance dilemma, *Neural Computation*, 1, pp. 1-50, 1992.

[57] P. Somol, J. Novovicova, Evaluating the stability of featurte selectors that optimize feature subset cardinality, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32, pp. 1-11, 2010.

[58] A. Kalousis, J. Prados, M.Hilario. Stability of feature selection algorithms: a study on high-dimensional spaces, *Knowledge and Information Systems*, 12, pp. 95-116, 2007.

[59] S. Stigler. Francis galton's account of the invention of correlation, *Statistical Science*,4(2), p. 7379, 1989.

[60] C. Spearman. The proof and measurement of association between two things, *The American journal of psychology*, 15, p. 72101, 1904.

[61] R Duda, P. Hart, D. Stork, *Pattern Classification*. *John Wiley & Sons*, New York (USA), 2001.

[62] R. Agrawal, R. Srikant. Fast algorithms for mining association rules. In *Proc. 20th International Conference on Very Large Data Bases (VLBD 1994)*,1, pp. 1-13, 1994.

[63] M. Sokolova, G. Lapalme. A systematic analysis of performance measures for classification tasks, *Information Processing and Management*, 45, pp. 427-437, 2009.

[64] S. Cateni, V. Colla, M. Vannucci. A method for resampling imbalanced datadata in binary classification tasks for real-world problems, *Neurocomputing*, 135, pp. 32-41, 2014.

[65] S. Cateni, V. Colla, M. Vannucci. Novel resampling method for the classification of imbalanced datasets for industrial and other real-world problems. In *International Conference on Intelligent Systems design and applications ISDA 2011*, pp. 402-407, 2011,

[66] A. Asuncion, D. Newman. UCI Machine Learning Repository - http://archive.ics.uci.edu/ml/datasets.html, 2007.

Author Biographies

Silvia Cateni (Cascina (PI), 11/03/1977) got her master degree in telecommunication engineering, in 2005, from University of Pisa. Her research activity includes mathematical modeling and data analysis through statistical and artificial intelligence-based techniques. She collaborated in several research projects particularly dealing with steelmaking industry. She is presently a research assistant at Scuola Superiore Sant'Anna.

Valentina Colla (La Spezia, 24/02/1970) received her master degree in telecommunication engineering from the University of Pisa in 1994 and her Ph.D. in Robotics from Scuola Superiore Sant'Anna in 1998. She became an associate professor at Scuola Superiore Sant'Anna in 2000 and She is presently a technical research manager at the Institute of Communication, Information and Perception Tecnologies (TeCIP) of Scuola Superiore Sant'Anna in Pisa. Her research fields include the application of neural networks and other Artificial Intelligence techniques to data analysis as well as to the simulation, monitoring and control of industrial processes and machineries. Valentina Colla is currently coordinator of the research center "ICT for Complex Industrial Systems and Processes (ICT-COISP)" of the TeCIP Institute.