A feature selection and classification algorithm based on randomized extraction of model populations

Aida Brankovic, Alessandro Falsone, Maria Prandini, *Member, IEEE,* and Luigi Piroddi, *Member, IEEE*

***Abstract*—We here introduce a novel classification approach adopted from the nonlinear model identification framework, which jointly addresses the feature selection and classifier design tasks. The classifier is constructed as a polynomial expansion of the original features and a selection process is applied to find the relevant model terms. The selection method progressively refines a probability distribution defined on the model structure space, by extracting sample models from the current distribution and using the aggregate information obtained from the evaluation of the population of models to reinforce the probability of extracting the most important terms. To reduce the initial search space, distance correlation filtering is optionally applied as a preprocessing technique. The proposed method is compared to other well-known feature selection and classification methods on standard benchmark problems. Besides the favorable properties of the method regarding classification accuracy, the obtained models have a simple structure, easily amenable to interpretation and analysis.**

***Index Terms*—Feature selection, Classification, Nonlinear identification, Model selection, Randomized methods.**

I. Introduction

In the supervised learning framework, classification is the task of predicting the class labels of unseen observations (each consisting of a set of measured attributes or features), based on the experience gathered through a learning process on a previously available set of observations whose classes are known (training set). The classification task is generally decomposed into two stages, namely a first preprocessing stage denoted feature selection (FS), followed by the actual classifier design. FS is a combinatorial optimization problem which aims at extracting the relevant features from a given set. An effective FS procedure greatly facilitates the classifier design process, reducing its computational demand, simplifying the classifier structure, and ultimately improving the classification performance, which may be adversely affected by irrelevant and redundant features [1]. FS is particularly crucial and hard in problems with a large number of features, resulting in a huge search space (curse of dimensionality).

FS methods are often classified according to how the attribute selection and model construction processes interact. In filter methods (see, *e.g.*, [2]), FS is performed independently of the classifier design, based only on intrinsic properties of the features, whereas in wrapper methods a subset of the features is evaluated based on the classification performance it can achieve if it is used to build the classifier. The classifier operates on the selected features processing them through a linear or nonlinear model. In this process it may be useful to derive additional more complex features as functions of the original ones (feature extraction).

Filter methods can be used effectively to eliminate the least important terms, but they cannot be fully relied upon to eliminate all redundant terms since they do not take into account the interaction between features. Such relationships between regressors may indeed have an important impact on the selection process. For example, individually important features may become redundant when considered in combination with others, and individually irrelevant or redundant features may become relevant in combination with others [3].

Wrapper methods are typically more accurate, though they are computationally intensive and may suffer from overfitting problems [4]. Several wrapper algorithms based on sequential search have been discussed in the literature, such as the PTA(l, r) (Plus l and Take Away r), the SFFS (Sequential Forward Floating Selection algorithm) and the SBFS (Sequential Backward Floating Selection algorithm) [5]. In these schemes, the algorithm starts from either the empty or the full set of features, and then features are iteratively added or removed. Similar incremental model building schemes have been developed in the context of nonlinear identification, particularly with reference to polynomial NARX/NARMAX models [6], [7], [8]. Besides sequential approaches, a significant amount of work has been devoted to evolutionary methods such as Genetic Algorithms (GA) [9], [10], Particle Swarm Optimization (PSO) [3], [11], Ant Colony Optimization (ACO) [12], [13], Harmony Search (HS) [14], Artificial Bee Colony (ABC) [15].

Other randomized approaches not based on evolutionary paradigms have been proposed in a few recent works. For example, the approach described in [16] involves an initial random selection of the feature subset, which is subsequently updated according to a randomized scheme that may substitute or remove a single feature with a given probability. Though the update process is randomized, the feature selection process is still incremental, with all the pros and cons of local search methods. Furthermore, the method only considers the performance of the current model as a whole, without distinguishing the relevance of the individual features within the model, thus making the update process blind with respect to the features. The methods discussed in [17] aim at the reduction of the

A. Brankovic, A. Falsone, M. Prandini, and L. Piroddi are with the Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Italy, e-mail: {aida.brankovic, alessandro.falsone, maria.prandini, luigi.piroddi}@polimi.it.

original search space by random projections of the feature space on a smaller space and random extractions of feature subsets on which to perform the classification task. Again, the strategy appears to be completely blind and does not exploit the information resulting from each single subset, that has been extracted and processed.

Regarding the classifier design problem, several algorithms have been proposed in the literature, based on Artificial Neural Networks (ANN) [18], Support Vector Machines (SVM), often in combination with Radial Basis Functions (RBF) as kernel functions [19], Twin Support Vector Machines (TSVM) [20], instance based learning methods such as Nearest Neighbor (NN) and Data Gravitational Classification (DGC) [21], evolutionary methods as genetic programming (GP) [22] and PSO [23], [24]. Instance based algorithms are particularly appealing due to their simple classification logic and generally satisfactory performance. In the NN (or 1-NN) algorithm, a new sample is simply assigned to the class of the nearest previously available sample. This is one of the most used and well known classification algorithms due to its simplicity, though it is reported to suffer from various drawbacks such as high dimensionality, low efficiency, high storage requirements and sensitivity to noise [25]. Cases where the classes are non-separable or overlapping appear to be particularly critical [26]. The k-NN extension classifies a new instance based on the majority of the k nearest neighbors. Several variants of the k-NN method have been proposed in the literature, that typically introduce weighted distances or similar concepts to improve the performance, such as the Adaptive k-NN (KNN-A) [27], the Distance Weighted k-NN (DW-KNN) [28], the Center NN (CNN) [29], the Cam weighted distance NN (CamNN) [30]. DGC algorithms [31] have also been put forward as an attempt to overcome the mentioned problems of the NN algorithm. In this respect, it is worth mentioning the work of Peng *et al.* [32], which employs feature weighting and a tentative random feature selection algorithm to compute the feature weights. An enhancement of the DCG algorithm, denoted DGC+ is proposed in [31] to deal with imbalanced data. Peng [33] also proposed a fast feature weighting algorithm for DGC that evaluates features by using discrimination and redundancy.

We here propose a novel wrapper algorithm, denoted in the following RFSC (Randomized Feature Selection and Classification), inspired by recent advancements in the context of nonlinear model identification [34]. Here, the mapping from features to classes – that is generally *nonlinear* – is constructed by a polynomial expansion of the original set of features and calculating the output as a linear combination of the extended features. This has the advantage that the resulting model configures a linear regression in the extended features (accordingly referred to as regressors), the parameters of which can be estimated by means of well-assessed techniques. In this framework, the selection of the most appropriate regressors to include in the model (denoted model structure selection or briefly MSS) constitutes the main challenge. The easiest way to address the MSS task is by incremental building procedures, akin to the sequential wrapper approaches mentioned previously. Such incremental approach is generally driven by an importance rating of the regressors, that measures the accuracy improvement that can be achieved by adding that particular term to the current model. Unfortunately, this is only a relative measure of the importance of a regressor, which may vary considerably depending on the other terms included in the model. Indeed, individually important regressors may become redundant in combination with others, while terms of scarce individual importance may become highly relevant in combination with others. This fact alone may greatly affect the correctness of the MSS process.

More reliable results can be obtained if the importance of the regressors is assessed based on information gathered from a *population* of models. In this perspective, along the lines of [34], we define a probability distribution over the model structure space, that describes the probability of each possible extended feature subset to be the true model structure. At each iteration, a population of sample model structures is extracted from the current distribution and all the corresponding models are estimated and evaluated. The aggregate information obtained by processing this population of models is used to update the probability distribution, by reinforcing the probability to extract those terms that appear in accurate models more often than not, and accordingly reducing the probability to extract the remaining ones. The method progressively refines the probability distribution until it converges to a limit distribution associated to a single model.

Experimental results show that this method provides quite compact and accurate models. The RFSC algorithm does not suffer from error accumulation problems that may be observed with sequential methods, and generally bases the selection of features on more robust evidence than what may be gathered from individual models. Also, the randomization inherent in the approach yields sufficient exploration capabilities to allow the algorithm to occasionally escape from local minima.

The search space rapidly increases with the number of features and the order of the polynomial expansion. Large-sized search spaces complicate the FS task and may adversely affect the search process. To address this issue, a dependency test based on the *correlation of distances* [35] has been carried out for medium/large size problems. Distance correlation provides a reliable dependency measure between random vectors, and can be used to test the individual dependence of the output vector on each feature vector. Only features with enough statistical evidence to reject the independence hypothesis with a given significance level are considered in the FS process, thus reducing the search space.

The rest of the paper is organized as follows. Section II presents the classification problem in a nonlinear regression framework. The proposed method is introduced in Section III. More in detail, a probabilistic formulation of the feature selection problem is provided first, which constitutes the basis for the development of the RFSC method, which is presented next. Finally, the introduction of a prefiltering method to reduce the feature space based on Distance Correlation Filtering concludes Section III. Several numerical studies on benchmark datasets are discussed in Section IV. Finally, Section V presents some concluding remarks.

II. PRELIMINARIES

A. The classification problem

In classification problems one is interested in constructing a model that captures the relationship between features (inputs) and classes (outputs) through a learning process operating on available observations (input-output pairs). Classification is akin to model identification, the main differences being that the input-output relationship is typically non-dynamic and that the outputs (and sometimes the inputs) take values in a discrete set. This similarity makes it sometimes possible to adapt algorithms developed in the identification domain to solve classification tasks, as endeavored here.

Assume that a set of N observations is available, each consisting of a pair $(\boldsymbol{u}(k), c(k))$, $k = 1, \ldots, N$, where the components u_p, $p = 1, \ldots, N_f$ of vector $\boldsymbol{u}$ are the features and $c \in \{1, \ldots, N_c\}$ is the corresponding class (assumed known, according to the supervised learning framework). In the following, we adopt a one-vs.-rest strategy to deal with multi-class problems, and accordingly recode the output as an N_c-dimensional vector $\boldsymbol{y}$, with binary components, defined as:

$$y_i(k) = \begin{cases} 1, & c(k) = i \\ -1, & \text{otherwise} \end{cases} \tag{1}$$

where $i = 1, \ldots, N_c$. Notice that if $N_c = 2$, a single output is sufficient to discriminate between the two classes, the -1 value of y_1 being directly associated to class 2.

The objective is to construct a classification model of the type:

$$\hat{\boldsymbol{y}}(k) = \boldsymbol{f}(\boldsymbol{u}(k)), \tag{2}$$

where $\hat{\boldsymbol{y}}$ denotes the class estimate and $\boldsymbol{f}$ is a vector of unknown functions, that is capable of predicting correctly the class for observations unseen in the learning phase. Following the one-vs.-rest strategy, a separate model is devoted to the assessment of each class.

To avoid ambiguities in the class estimation, the actual class estimate is conventionally assumed as the label corresponding to the individual classifier returning the largest value:

$$\hat{c}(k) = \arg \max_{i=1,\ldots,N_c} \hat{y}_i(k). \tag{3}$$

In view of this, the multi-class problem can be addressed by training one binary classifier for each class, that discriminates if a sample belongs to one class or not. Accordingly, in the following we shall focus on the training and evaluation of the binary classifiers $\hat{y}_i(k)$, $i = 1, \ldots, N_c$.

Regarding the unknown functions $f_i(\cdot)$, $i = 1, \ldots, N_c$, a common approach is to represent them using parametric functional expansions, so that:

$$\hat{y}_i(k) = \left(\sum_{j=1}^{N_r} \vartheta_j^{(i)} \varphi_j(k) \right) = \Phi(k)^T \vartheta^{(i)}, \tag{4}$$

$i = 1, \ldots, N_c$, where $\vartheta^{(i)}$ is a vector of unknown parameters (associated to the ith output), and $\Phi(k) = [\varphi_1(k) \ldots \varphi_{N_r}(k)]^T$, where $\varphi_j(k) = \varphi_j(\boldsymbol{u}(k))$, $j = 1, \ldots, N_r$, is a given nonlinear function of the features. In view of the fact that equation (4) actually configures a linear regression, these extended features are also called regressors.

Various types of basis functions have been used to construct the functional expansions, such as Fourier series, piecewise linear models, polynomial models, radial basis functions, and sigmoidal neural networks, all having the universal approximation property. In this work, we will consider polynomial expansions, so that the generic term $\varphi_j(k)$ takes the form:

$$\varphi_j(k) = \begin{cases} u_{p_1}(k) \cdot u_{p_2}(k) \cdot \ldots \cdot u_{p_l}(k), & l > 0 \\ 1, & l = 0 \end{cases} \tag{5}$$

where $p_s \in \{1, \ldots, N_f\}$, $s = 1, \ldots, l$, with $0 \leq l \leq M$, M being the maximum allowed degree of the polynomial expansion.

This formulation has the advantage that the model is linear-in-the-parameters, which greatly facilitates parameter estimation. On the other hand, the number of terms in a polynomial expansion increases rapidly with the maximum degree and the number of arguments (curse of dimensionality). Conventional practice has it that relatively small models of this category are suitable for various applications. It is also well-known that the smaller the size of the model, the more robust it will be and the more capable of generalizing to new observations. Therefore, a crucial problem consists in selecting the best terms of type (5) for the model, a task which is equivalent to feature selection, but applied to an extended set of features (constructed as monomials of the original ones).

B. Parameter estimation of the ith component of the classifier

As already mentioned, the modeling task is addressed separately for each class. In the following, we shall focus on the modeling of classifier $\hat{y}_i$ associated to class i. For ease of notation we will drop the indexing on class i.

For a given structure, the parameter estimation for a model of type (4) is typically formulated as a minimization problem with reference to a loss function defined as $\mathcal{L} : \{-1, +1\} \times \mathbb{R} \to \mathbb{R}_+$. A standard loss function evaluates the model performance as the percentage of misclassified observations (with respect to class i). The resulting optimization problem is given by

$$\min_{\vartheta} \ \frac{1}{N} \sum_{k=1}^{N} \mathcal{L}_{0-1}(y(k), \hat{y}(k)), \tag{6}$$

where $\mathcal{L}_{0-1}$ is the *0-1 loss function*, defined as $\mathcal{L}_{0-1}(z_1, z_2) = \mathbf{1}_{\{z_1 z_2 < 0\}}(z_1, z_2)$. Due to the hard nonlinearity enforced by this loss function, the latter is usually approximated with functions with nicer properties regarding optimization (*e.g.*, hinge, squared hinge, logistic, exponential). In the following, the *logistic loss* will be employed for this purpose, resulting in the following reformulation of the optimization problem:

$$\min_{\vartheta} \ \frac{1}{N} \sum_{k=1}^{N} \log(1 + e^{-y(k)\hat{y}(k)}). \tag{7}$$

The reader should note that $\log(1 + e^{(\cdot)})$ is a strictly convex function, and $\hat{y}(k)$ is linear in ϑ. Therefore, the resulting cost

function is strictly convex in ϑ, and the minimizer of (7) is unique.

Although a closed-form solution to the above optimization problem does not exist, the logistic loss is a continuous differentiable function, which allows to apply gradient descent methods in the optimization process. In this work, a standard Newton's iterative optimization scheme is applied to solve problem (7).

C. Statistical test for regressor significance

The rejection of redundant terms is a crucial step in the identification procedure. For this purpose, a statistical test is carried out after the parameter estimation phase to rule out terms whose parameters are statistically indistinguishable from 0. According to [36], the update equation of the Newton method is structurally equivalent to an Iteratively Reweighted Least Squares algorithm, so that upon convergence one can evaluate the parameter covariance as in standard Weighted Least Squares. Therefore, the variance $\hat{\sigma}_j^2$ of the estimated parameters is given by:

$$\hat{\sigma}_j^2 \approx \hat{\sigma}_e^2 G_{jj}^{-1}, \tag{8}$$

where $\hat{\sigma}_e^2$ is variance of the residuals and G_{jj} is the jth diagonal element of the Hessian $G = \Psi^T R \Psi$ upon convergence, $\Psi = [\Phi(1) \dots \Phi(N)]^T$ being a matrix containing all samples of the selected nonlinear regressors and R a diagonal $N \times N$ matrix with diagonal elements given by:

$$R_{kk} = \tilde{y}(k)(1 - \tilde{y}(k)), \tag{9}$$

$k = 1, \dots, N$, where $\tilde{y}(k) = 1/(1 + e^{y(k)\hat{y}(k)})$.

The variance (8) can be employed in a Student's t-test to determine the statistical relevance of each regressor [34]. More precisely, the jth regressor is considered to be statistically irrelevant (and therefore rejected) if the interval

$$[\hat{\vartheta}_j - \hat{\sigma}_j t_{\alpha, N-\tau}, \hat{\vartheta}_j + \hat{\sigma}_j t_{\alpha, N-\tau}] \tag{10}$$

contains 0, $t_{\alpha,N-\tau}$ being the $100(1-\alpha)$ percentile of the Student's t distribution with $N - \tau$ degrees of freedom and significance confidence interval α, where τ is the number of components of ϑ.

III. The proposed method

A. Probabilistic formulation of the Feature Selection problem

This section addresses the FS problem for the ith component of the classifier. As done in the previous section, we drop the indexing on class i, for simplicity.

FS can be envisaged as the problem of finding the subset of regressors that maximizes the performance index J:

$$J(f) = 1 - \frac{1}{N} \sum_{k=1}^{N} \mathcal{L}_{0-1}(y(k), \hat{y}(f; k)) \tag{11}$$

over all possible model structures f in the set $\mathcal{F} = 2^{\mathcal{R}}$, $\mathcal{R} = \{\varphi_1, \cdots, \varphi_{N_r}\}$ being the full set of N_r regressors. Notice that the dependence of the classifier on the model structure f has been explicitly stated in Equation (11).

This problem is combinatorial, in that in principle one would need to explore all 2^{N_r} model structures. A convenient solution approach involves a reformulation in a probabilistic framework [34] by introducing the random variable ϕ which takes values in the set of all possible models $\mathcal{F}$ according to a probability distribution $\mathcal{P}_\phi$. The performance of ϕ is also a random variable, and its expectation is given by

$$\mathbb{E}[J(\phi)] = \sum_{f \in \mathcal{F}} J(f) \mathcal{P}_\phi(f). \tag{12}$$

Index (12) is maximized when the probability mass concentrates on the model structure associated to the highest value of J (or one of the possible best model structures, if the minimum is not unique).

Therefore, the problem of finding the best $f \in \mathcal{F}$ can be formulated as

$$\mathcal{P}_\phi^* = \arg \max_{\mathcal{P}_\phi} \mathbb{E}[J(\phi)], \tag{13}$$

where $\mathcal{P}_\phi^*$ is such that $\mathcal{P}_\phi^*(f^*) = 1$.

A convenient parametrization for $\mathcal{P}_\phi$ is obtained by associating a Bernoulli random variable ρ_j to each regressor φ_j, that models the probability that φ_j belongs to the target model:

$$\rho_j \sim Be(\mu_j), \tag{14}$$

$j = 1, \dots, N_r$, where $\mu_j \in [0, 1]$. According to this representation, a model extraction from $\mathcal{P}_\phi$ implies testing each regressor for inclusion, by extracting a value from the respective Bernoullian distribution. Regressor φ_j is included if the outcome of the jth extraction is 1, and omitted in case of 0. The former event has probability μ_j, whereas the probability of getting a 0 is given by $1-\mu_j$. Accordingly, in the rest of the paper we will denote μ_j as the *Regressor Inclusion Probability* (RIP) of the jth regressor, and define $\mu = [\mu_1 \cdots \mu_{N_r}]^T$ as the vector of RIPs. For simplicity, we assume that all random variables ρ_j, $j = 1, \dots, N_r$ are independent. In summary, the probability distribution $\mathcal{P}_\phi$ over the models in $\mathcal{F}$ can be written as:

$$\mathcal{P}_\phi(f) = \prod_{j:\varphi_j \in f} \mu_j \prod_{j:\varphi_j \notin f} (1 - \mu_j) \tag{15}$$

for any $f \in \mathcal{F}$. If all RIPs have values in $\{0, 1\}$ only, a limit distribution is obtained with all probability mass concentrated on a specific model $\tilde{f}$ (containing all the regressors whose RIPs equal 1). In that case, it follows that $\mathcal{P}_\phi(\tilde{f}) = 1$. The objective of the FS procedure will therefore be that of adapting the RIPs until convergence to the target limit distribution associated to an optimal model f^*.

To evaluate the importance of a given term we consider an aggregate indicator $\mathcal{I}_j$ that compares the average performance of the models including the jth regressor with that of the remaining ones:

$$\mathcal{I}_j = \mathbb{E}[J(\phi) | \varphi_j \in \phi] - \mathbb{E}[J(\phi) | \varphi_j \notin \phi], \tag{16}$$

where $j = 1, \dots, N_r$. The interested reader is referred to [34] for all the mathematical details. Thanks to the averaging over all models, indicator $\mathcal{I}_j$ can be interpreted as a global measure of the regressor importance. In [34], the authors prove that if $\mathcal{P}_\phi(f^*)$ is sufficiently high, then $\mathcal{I}_j$ takes positive values when $\varphi_j \in f^*$ and negative otherwise.

B. The Randomized Feature Selection and Classification algorithm

In view of the probabilistic reformulation of the FS problem discussed in the previous section, we here describe an iterative optimization approach that operates on the model distribution $\mathcal{P}_\phi(f)$ with the aim of maximizing the average performance given by (12). In detail, the RIPs are progressively updated based on the assessment of the importance of each regressor in terms of index $\mathcal{I}_j$, $j = 1, \cdots, N_r$. Notice that an exact evaluation of $\mathcal{I}_j$ is not practically feasible, since it would require an exhaustive approach on the model space. Therefore, the expected values in (16) are approximated with averages over a finite set of models extracted from the current model distribution. The procedure is stopped upon reaching a limit distribution.

Given the discrete nature of the $0-1$ loss function in FS problems, different models may result in the same classifier or in different classifiers of equal performance, unlike what happens in MSS in the nonlinear identification framework. Therefore, it may happen that different runs of the algorithm may provide different results.

At the beginning of each iteration, a set of models is extracted from the space of all possible model structures using the current Bernoullian distributions associated to the regressors. More in detail, for each model a value is extracted from all distributions, and only the regressors corresponding to a successful extraction are included in the model. Then, the parameters of each model are estimated and its performance evaluated according to the procedure explained in II-B (if any redundant terms are detected, they are eliminated and the parameters re-estimated prior to evaluation). Then, the following update law is applied to the Bernoullian distribution of each regressor at the tth iteration:

$$\mu_j(t+1) = \text{sat}(\mu_j(t) + \gamma \tilde{\mathcal{I}}_j) \tag{17}$$

for $j = 1, \cdots, N_r$, where $\tilde{\mathcal{I}}_j$ is an approximation of $\mathcal{I}_j$ calculated on the set of extracted models, $\text{sat}(x) = \min(\max(x, 0), 1)$ is a function that ensures that the calculated μ_j values will not exceed the interval $[0, 1]$, and γ is a step-size parameter. The value of the latter parameter is adapted at each iteration:

$$\gamma = \frac{1}{10(J_{max} - \bar{J}) + 0.1}, \tag{18}$$

where J_{max} is the performance index of the best model among those extracted at the current iteration and $\bar{J}$ is their average performance. In practice, the larger the variance of the model performances, the smaller the step-size, indicating a lower level of reliability of the computed global measure of the regressor importance $\mathcal{I}_j$.

The procedure is iterated as long as the RIPs continue to be modified.

A sketch of the basic loop of the proposed RFSC procedure is presented below as Algorithm 1.

Algorithm 1 Main loop of the RFSC.

Input: $\{\boldsymbol{u}(k), \boldsymbol{y}(k)\}$, $\mathcal{R} = \{\varphi_1, \ldots, \varphi_{N_r}\}$, N_p, α, $\mu(0)$, ϵ, N_i

Output: μ

```
for i = 1 to N_i do
    for n_p = 1 to N_p do
        Generate random model structure ∈ F
        Estimate parameter vector ϑ by solving (6)
        Compute σ̂_j^2 according to (8)
        Apply statistical test according to (10)
        Remove redundant terms
        Re-estimate parameter vector ϑ
        Evaluate model performance according to (11)
    end for
    for j = 1 to N_r do
        Evaluate importance of jth term using (16)
        Update jth RIP according to (17)
    end for
    t ← t + 1
    if max_{j=1,...,N_r} |μ_j(t) − μ_j(t−1)| ≤ ε then
        Break
    end if
end for
```

C. Prefiltering of the feature space using Distance Correlation Filtering

A high-dimensional feature space can hamper FS algorithms by slowing down the search process and by increasing the chances of getting stuck in local minima. To tackle this issue a common approach is to perform a prefiltering of the feature space. Specifically, it would be desirable to identify those features that are relevant in describing the output, and those which are not. We address this problem by analyzing the dependence of the output on each feature, according to the rationale that if feature u_p is not important in the description of the output y_i, then we would expect y_i and u_p to be independent. The reader should note that at this point we are just interested in characterizing the dependence/independence of the output from a specific feature, not the strength nor the "shape" of such dependence, tasks that are performed during the FS process.

There exist various statistical tests designed to assess the dependence between two random vectors. We here employ the one described in [35], which is based on a statistic named "distance correlation". The statistical test in [35] is very flexible and can handle both discrete and continuous random vectors, without any assumption on their distributions, making it particularly amenable for classification purposes. For the sake of completeness, we here briefly report the main result of [35].

Let X and Y be two random variables such that $\mathbb{E}[|X| + |Y|] < \infty$, where $|\cdot|$ denotes the absolute value. In our case we have $X = u_p$ and $Y = y_i$ for any i and p. We want to test the null hypothesis $H_0 : X$ *and* Y *independent*. Let $\boldsymbol{X} = [u_p(1) \cdots u_p(N)]^T$ be a vector of i.i.d. realizations of X, and $\boldsymbol{Y} = [y_i(1) \cdots y_i(N)]^T$ the corresponding realizations of Y.

Now define the "empirical distance covariance" as

$$\nu_N^2(\boldsymbol{X},\boldsymbol{Y}) = \frac{1}{N^2}\sum_{r,s=1}^{n} A_{r,s}B_{r,s}, \tag{19}$$

where $A_{rs} = a_{rs} - \bar{a}_{r\cdot} - \bar{a}_{\cdot s} + \bar{a}_{\cdot\cdot}$, $B_{rs} = b_{rs} - \bar{b}_{r\cdot} - \bar{b}_{\cdot s} + \bar{b}_{\cdot\cdot}$, with $a_{rs} = |u_p(r) - u_p(s)|$, $b_{rs} = |y_i(r) - y_i(s)|$, and $\bar{a}_{r\cdot} = \frac{1}{N}\sum_{s=1}^{N} a_{rs}$, $\bar{a}_{\cdot s} = \frac{1}{N}\sum_{r=1}^{N} a_{rs}$, $\bar{a}_{\cdot\cdot} = \frac{1}{N^2}\sum_{r,s=1}^{N} a_{rs}$, $\bar{b}_{r\cdot} = \frac{1}{N}\sum_{s=1}^{N} b_{rs}$, $\bar{b}_{\cdot s} = \frac{1}{N}\sum_{r=1}^{N} b_{rs}$, $\bar{b}_{\cdot\cdot} = \frac{1}{N^2}\sum_{r,s=1}^{N} b_{rs}$.

The interested reader is referred to [35] for detailed information on $\nu_N(\boldsymbol{X},\boldsymbol{Y})$ and its properties.

The statistical test proposed in [35] rejects H_0 if

$$\frac{N\,\nu_N^2(\boldsymbol{X},\boldsymbol{Y})}{S} > \mathcal{N}^{-1}\Big(1-\frac{\alpha_d}{2}\Big)^2, \tag{20}$$

where $\mathcal{N}(\cdot)$ denotes the normal cumulative distribution function, α_d is the significance level of the test, and

$$S = \bar{a}_{\cdot\cdot}\bar{b}_{\cdot\cdot}. \tag{21}$$

For each i, inequality (20) is tested for all p, and only those features u_p for which there is enough statistical evidence to reject the independence hypothesis are considered in the FS process for determining the classifier $\hat{y}_i$. The prefiltering procedure, denoted distance correlation filtering (DCF), is summarized in Algorithm 2.

Algorithm 2 Feature set preprocessing for class i.

Input: $\{\boldsymbol{u}(k), y_i(k)\}$, $\mathcal{F}_s = \{u_1,\ldots,u_{N_f}\}$, α_d
Output: $\tilde{\mathcal{F}}_s^i$

```
F̃_s^i ← F_s
for j = 1 to N_f do
    H_0^j ← true
    X ← [u_j(1) ··· u_j(N)]^T
    Y ← [y_i(1) ··· y_i(N)]^T
    Compute ν_N^2(X, Y) as in (19)
    Compute S as in (21)
    if N ν_N^2(X, Y)/S > N^{-1}(1 − α_d/2)^2 then
        H_0^j ← false
    end if
    if H_0^j then
        F̃_s^i ← F̃_s^i \ {u_j}
    end if
end for
```

IV. Experimental study

A. Experiment design and datasets

This section illustrates various experiments carried out to assess the performance of the proposed algorithm. Eight numerical datasets from the UCI machine learning repository [37] have been analyzed. The main features of the selected datasets are given in Table I.

All the input data in the original feature sets have been normalized in the range $[0,1]$ range according to:

$$u_p(k) = \frac{u_{p,raw}(k) - u_{p_{min}}}{u_{p_{max}} - u_{p_{min}}}, \tag{22}$$

for $k = 1,\ldots,N$, where $u_{p,raw}(k)$ is the original numeric value of the kth observation of feature p in a given dataset, and $u_{p_{max}}$ and $u_{p_{min}}$ represent the maximum and minimum value of the pth attribute in the dataset, respectively.

The classification performance of the proposed algorithm on the selected datasets has been evaluated using the 10-fold cross validation (10-FCV) approach. Briefly, the dataset is split into ten (equal and non-overlapping) subsets (folds), possibly uniformly representative of all classes. Nine folds are used for training and the remaining one for testing, the procedure being repeated 10 times so that all folds are tested once. The algorithm performance is finally computed as the average over the ten independent runs. Given the randomized nature of the RFSC, different results may be obtained on each run, especially on datasets with large feature sets, for which full exploration may be too costly. For this reason, the application of the RFSC on each fold is repeated 10 times and the best model retained. A different criterion has been adopted for the HillValley and Musk1 datasets, to obtain results more directly comparable to the literature. Specifically, 70% samples are employed for training and 30% for testing. The algorithm performance is computed as the average of 10 independent runs with a random data division of the training-testing pairs.

The classifier performance can be evaluated in terms of the percentage of correct classifications. In addition, we provide an alternative accuracy measure, namely the Cohen's Kappa rate [38], which is capable of dealing more effectively with imbalanced data.

The Kappa statistic was originally designed to compare two different classifiers to measure the degree of (dis)agreement, compensating for chance (dis)agreements, but can be used to evaluate the merit of a specific classifier by comparing it to an "ideal" classifier producing the exact classifications. Let the *confusion matrix* be an $N_c \times N_c$ matrix C such that C_{ij} equals the number of samples that are classified in class i by classifier 1 and j by classifier 2, and denote by $C_{i\cdot} = \sum_{k=1}^{N_c} C_{ik}$ and $C_{\cdot j} = \sum_{k=1}^{N_c} C_{kj}$ the row and column counts (that represent the individual classification counts). Then, the Kappa rate is defined as follows:

$$K = \frac{N\sum_{i=1}^{N_c} C_{ii} - \sum_{i=1}^{N_c} C_{i\cdot}C_{\cdot i}}{N^2 - \sum_{i=1}^{N_c} C_{i\cdot}C_{\cdot i}}, \tag{23}$$

and ranges from -1 (total disagreement) to 0 (random classification) to 1 (total agreement). The Kappa statistic is very useful for multi-class problems, in that it measures the classifier accuracy while compensating for random successes [31].

TABLE I
Main characteristics of the used datasets, [37].

Dataset name	No. of samples	No. of features	Feature types Real	Feature types Integer	No. of classes
Bupa	345	6	1	5	2
HillValley	606	0	100	0	2
Ionosphere	351	34	32	1	2
Iris	150	4	4	0	3
Musk1	476	166	0	166	2
Sonar	208	60	60	0	2
WDBC	569	30	13	0	2
Wine	178	13	13	0	3

Regarding the initial parameter setup for the RFSC, the number of iterations was set to $N_i = 300$, the maximum nonlinearity degree to $N_d = 2$, the number of generated models to $N_p = 100$, the significance confidence interval to $\alpha = 0.99$ and all initial RIPs to $\mu_0 = 1/N_r$. Parameter α also influences the average model size, by acting on the threshold for the rejection of redundant terms. The closer α is to 1, the more regressors are rejected (and greater is the confidence that only meaningful regressors are retained), and the smaller is the average model size. Parameter ϵ in the termination condition has been set to $\epsilon = 0.002$. Finally, the significance level α_d for the prefiltering phase was set to 0.99 for the HillValley, Ionosphere, and Musk1 databases, to 0.87 for the Sonar database, and to 0.9999 for the WBCD database. The proposed algorithm was implemented in Matlab (version 2012b) and executed on an Intel(R) Core i7-3630QM machine, with 2.4GHz CPU, 8GB of RAM, and a 64-bit Operating System.

B. An illustration example

To get a greater insight in the mechanisms of the selection process, we here illustrate the RFSC behavior with reference to the WDBC dataset, which has 30 attributes and 2 class labels. Assuming a maximum nonlinearity degree of $N_d = 2$, the total number of extended regressors is $N_r = 496$. We will focus on two independent runs of the RFSC algorithm. Both runs returned a 7-terms model (denoted Model 1 and Model 2) with no common regressors and only one common feature. We refer to the regressors of the returned models as "final" regressors. It is worth mentioning that despite their different structure, Model 1 and Model 2 both exhibit 0 classification errors on the validation dataset.

Figures 1-2 (top) show the evolution of the RIP values for both runs. In both cases various regressors are initially considered promising and their RIPs increased. In the first run (Fig. 1, top) the RIPs of the final regressors keep increasing from the very first iterations and the other regressors are progressively discarded as the algorithm progresses. On the other hand, in the second run (Fig. 2, top) most regressors are selected or discarded in the first 25 iterations, but the last regressor is selected at a later stage (around iteration 40), essentially after two other terms have been rejected. Before final convergence, other regressors are tested but ultimately discarded. It is interesting to note that in both cases some regressors are initially selected, to the point that their RIPs rise to 1, but are subsequently rejected in favor of other terms. If we compare (column-wise) this behavior of the RIPs with the evolution of the average loss function (average value of the loss function of the N_p extracted models at a given iteration) in Figures 1-2 (middle), it is clear that the algorithm is exploring model structures with a higher average loss function in order to ultimately escape from structures that represent only local minima.

Figures 1-2 (bottom) show the average model size (AMS) at each iteration for both runs. For Model 1, the AMS of the generated models (measured before the application of the statistical test) grows rapidly in the beginning and starts

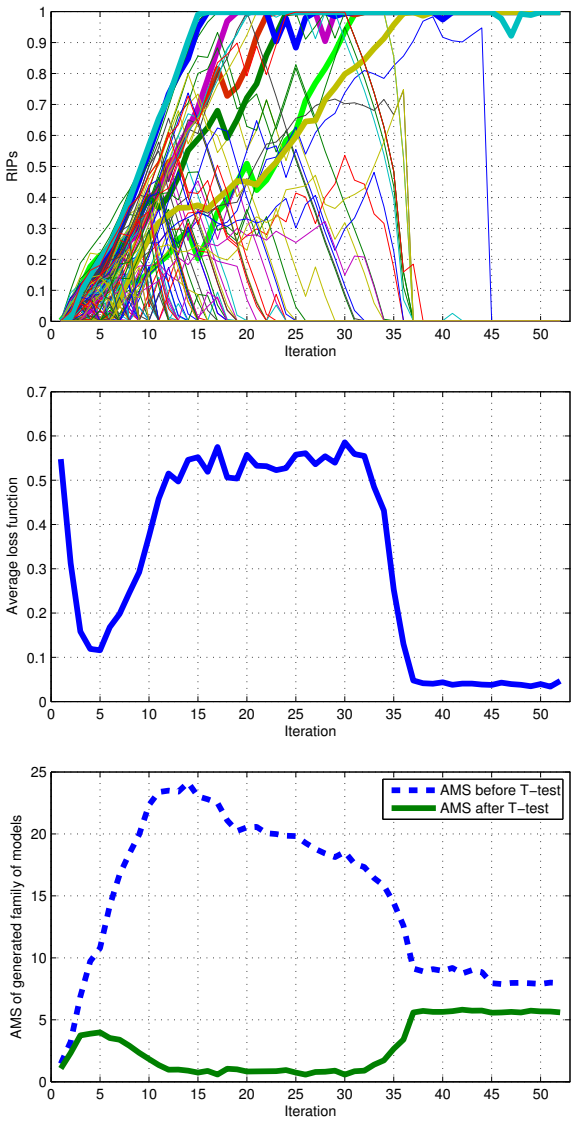

Fig. 1. Model 1: Evolution of the RIPs during the selection process (top, thicker lines indicate the terms contained in the final model), average loss function (middle), average model size (bottom) before (dashed) and after (solid) the t-test.

decreasing significantly only after iteration 10. Later on, after iteration 38, the model size does not change significantly. On the other hand, the AMS measured after the statistical test is very low from iterations 10 to 30, indicating that the algorithm is systematically rejecting tentative regressors as redundant. It is only between iterations 30 to 40 (*i.e.*, when the final two regressors have been added), that the model size converges to its final value. Similarly, for Model 2 the AMS before the t-test increases at the beginning, reaching a peak around iteration 15, and then it stabilizes after iteration 20.

Notice that in both runs the AMS value is always reduced after the test, indicating the effectiveness of the latter in detecting redundant terms.

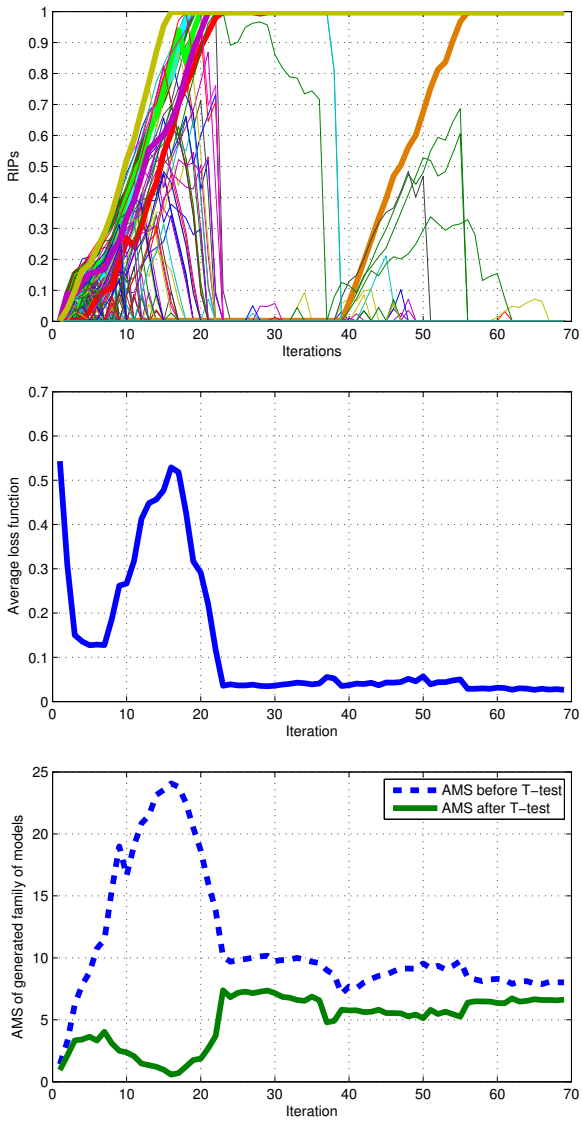

Fig. 2. Model 2: Evolution of the RIPs during the selection process (top, thicker lines indicate the terms contained in the final model), average loss function (middle), average model size (bottom) before (dashed) and after (solid) the t-test.

C. Interpretation of the results

As previously stated, all input data points $u_p(k)$, with $p = 1, \ldots, N_f$ and $k = 1, \ldots, N$, have been normalized to be in the $[0, 1]$ interval. Since each regressor $\varphi_j(k)$ is constructed as a product of features, $\varphi_j(k)$ takes values in $[0, 1]$ as well, for all $j = 1, \ldots, N_r$ and $k = 1, \ldots, N$.

Now, the estimated model is of the form (4), where only the selected regressors are associated to non-zero parameters. The predicted class for the kth observation is given only by the sign of $\hat{y}_i$, while the absolute value of $\hat{y}_i$ is related to the reliability of the prediction. Since $\varphi_j(k)$ is non-negative, the information about the sign is carried by the coefficients $\vartheta^{(i)}$ of the linear combination in (4). Therefore, the model can be decomposed in two additive components based simply on the sign of the parameters:

$$\hat{y}_i(k) = \hat{y}_i^+(k) - \hat{y}_i^-(k) = \Phi(k)_+^T \vartheta_+^{(i)} - \Phi(k)_-^T(-\vartheta_-^{(i)}), \quad (24)$$

where the first component $\hat{y}_i^+(k) = \Phi(k)_+^T \vartheta_+^{(i)}$ is associated to terms with positive coefficients and the second one $\hat{y}_i^-(k) = \Phi(k)_-^T(-\vartheta_-^{(i)})$ to terms with negative coefficients. This decomposition has the following very nice and clear interpretation: features which appear in regressors inside $\hat{y}_i^+(k)$ are representative for class i, whereas features appearing in $\hat{y}_i^-(k)$ are against class i. The "strongest" group of (extended) features in the ith model determines the sign of $\hat{y}_i$, and therefore if the predicted class should be class i or not. If multiple classes exhibit a positive $\hat{y}_i$, then the class is determined by the most "confident" classifier, *i.e* the one with the largest difference between $\hat{y}_i^+(k)$ and $\hat{y}_i^-(k)$.

In Figure 3, we report the values of the two quantities $\hat{y}_i^+(k)$ and $\hat{y}_i^-(k)$ for 20 validation data points. The two plots in Figure 3 (top and bottom) refer to the final models of the two runs of the RFSC algorithm analyzed in the previous section. Both models exhibit 0 classification errors on the validation set (56 samples).

From Figure 3, it is also apparent that despite the fact that both models exhibit 0 classification errors, they are not equivalent in terms of reliability. In particular, the value of $\delta_i(k) = (\hat{y}_i^+(k) - \hat{y}_i^-(k))/\max(\hat{y}_i^+(k), \hat{y}_i^-(k))$ can be interpreted as the "confidence" the model has in attributing class i to the kth sample. Apparently, Model 1 has generally greater values of δ_i. This difference is not currently captured by the performance index (11), and therefore the two models are considered equivalent for the RFSC algorithm.

To conclude the analysis of the results, we report in Table II the average size of the final model structures obtained by the 10-FCV procedure. Specifically, Table II displays the number of original attributes N_a, the number of attributes after the DCF procedure N_a^*, the average number of attributes n_a used by the classifier over the 10 folds, the number of regressors N_r generated based on the original attributes, the number of regressors N_r^* generated based on the filtered attributes, the average number of regressors n_r used by the classifier over the 10 folds. In the non-binary classification problems (Iris and Wine datasets), a separate modeling is carried out for each class. In those cases, the classifier size (in terms of number of used features and regressors) is calculated by performing the union over the individual class models $\hat{y}_i$, $i = 1, \ldots, N_c$.

By inspecting Table II, it is noticeable that while the RFSC algorithm employs a considerable fraction of the available features, it generally requires only a small number of regressors, demonstrating its capability of compressing the information in few terms.

D. Comparative analysis

To assess the performance of the RFSC algorithm, we report in this section an extensive comparison with the results documented in [3], [13], [15], [31], [33], [39], on the datasets in Table I. The comparison is carried out in terms of the average classification accuracy J_a, the average Kappa rate K_a, and the average model size. The performance comparison

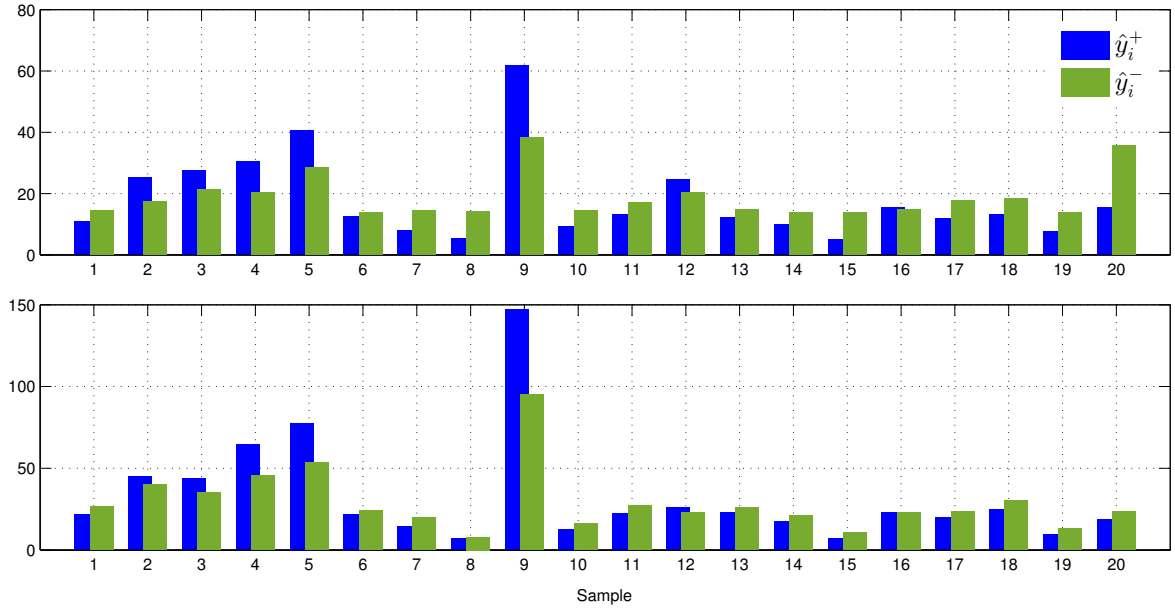

Fig. 3. Values of $\hat{y}_i^+(k)$ and $\hat{y}_i^-(k)$ for 20 validation samples: Model 1 (top) and Model 2 (bottom).

is summarized in Tables III-IV and the size comparison in Table V. The best result within a column is highlighted.

The RFSC outperformed all other documented results on the Bupa, HillValley and WDBC datasets, both in terms of average accuracy and average Kappa rate. This has been achieved at the cost of using more attributes compared to the other methods. On the other hand, the proposed algorithm was only slightly outperformed by the best competitor (which is different from case to case) on the remaining datasets, providing overall a good tradeoff between model complexity and accuracy. Considering more recent approaches such as [13], [15], and [33], the proposed RFSC dominates on 4 out of 8 tested benchmarks, and is only slightly outperformed in the other cases.

E. Computational time

A comparative analysis in terms of computational time is finally presented in Table VI. Though inherently time consuming due to model exploration mechanism in the randomized MSS process, the RFSC achieves convergence in comparable time with competitor algorithms. Indeed, it outperforms the PSO4-2 method for the Wine and WDBC datasets, but is generally somewhat slower than PSO+LDA. In this respect, it is important to note that non-optimized Matlab code has been used to obtain the documented results, so that the reported figures must be considered gross upper bounds. Still, the computational time goes from a few seconds to a little more than a minute for all analyzed datasets.

TABLE II
AVERAGE SIZE OF THE OBTAINED CLASSIFIERS OVER THE 10 FOLDS.

Dataset	N_a	N_a^*	n_a	N_r	N_r^*	n_r
Bupa	6	–	5.8	28	–	7.4
HillValley	100	100	8.3	5151	5151	3.7
Ionosphere	34	29	16.4	595	465	14.7
Iris	4	–	3.2	15	–	6.1
Musk1	166	165	46.2	14028	13860	23.2
Sonar	60	39	25.8	1891	820	18.7
WDBC	30	24	11.5	496	325	10.3
Wine	13	–	7.3	105	–	7.5

In order to analyze the algorithm's computational effort as a function of the problem size, we report in Fig. 4 the elapsed time versus the number of extended features for the WDBC dataset. More precisely, the curves in Fig. 4 show, for different values of the maximum number of iterations N_i, the average computational time associated to the algorithm calculated over ten different subsets of extended features of a given size (drawn at random). The four curves are characterized by an initial increase of the computational time with the problem size, followed by a saturation. The latter indicates that all simulations exhaust the available number of iterations above a certain problem size. Notice that the initial RIPs are defined so as to result in the same initial AMS for any problem size, so that it is expected that the computational load of the algorithm is essentially independent of the model size for a given N_i. As N_i increases, the saturation point shifts, indicating that early convergence is sometimes achieved.

V. CONCLUSION

A novel method has been proposed to jointly address the FS and classifier design problems, inspired by recent results in the nonlinear model identification domain. The FS problem is reformulated as a model structure selection problem where suitable nonlinear functions of the original features are evaluated for insertion in a linear regression model. Differently from commonly adopted methods, the importance of each candidate regressors is not evaluated with reference to a specific model, but to an ensemble of models, which appears to provide a more reliable information regarding the actual significance of the term. A distribution of models is used to extract the ensemble of models and is then updated based on the aggregate

TABLE III
COMPARATIVE PERFORMANCE ANALYSIS IN TERMS OF ACCURACY.

FS Method + Classifier	Ref.	Bupa	HillValley	Ionosphere	Iris	Musk1	Sonar	WBCD	Wine
ACO + PMC	[13]	0.6725	-	0.9373	0.9600	-	0.9087	-	0.9755
Att. WV + DGC	[31]	0.6744	-	0.9311	0.9533	-	0.8487	0.9619	0.9731
Att. WV + DGC	[31]	0.6525		0.6724	0.9533	-	0.7694	-	0.9706
- + KNN	[31]	0.6066	-	0.8519	0.9400	-	0.8307	-	0.9549
- + KNN-A	[31]	0.6257	-	0.9372	0.9533	-	0.8798	-	0.9663
- + DW-KNN	[31]	0.6376	-	0.8747	0.9400	-	0.8648	-	0.9438
- + Cam-NN	[31]	0.5962	-	0.7379	0.9467	-	0.7743	-	0.9497
- + CNN	[31]	0.6316	-	0.8917	0.9267	-	0.8940	-	0.9663
SSMA + SFLDS	[31]	0.6426	-	0.9088	0.9533	-	0.8079	-	0.9438
forward FS + LDA	[39]	0.6110	-	0.8530	0.9630	-	0.7610	-	0.9660
backward FS + LDA	[39]	0.6430	-	0.9090	0.9370	-	0.8550	-	0.9990
PSO + LDA	[39]	0.6520	-	0.9220	0.9700	-	0.9050	-	1.000
PSO(4-2) + 5NN	[3]	-	0.5777	0.8727	-	0.8494	0.7816	0.9398	0.9526
FFW-DGC	[33]	-	-	0.9461	0.9667	-	0.9173	0.9525	0.9831
MDis ABC	[15]	-	0.5508	-	-	0.8529	-	-	-
DCF + RFSC		0.7800	0.9277	0.9330	0.9666	0.8132	0.8806	0.9827	0.9944

TABLE IV
COMPARATIVE PERFORMANCE ANALYSIS IN TERMS OF KAPPA RATE.

FS Method + Classifier	Ref.	Bupa	HillValley	Ionosphere	Iris	Musk1	Sonar	WBCD	Wine
ACO + PMC	[13]	0.3259	-	0.8604	0.9400	-	0.8164	-	0.9659
Att. WV + DGC	[31]	0.3076	-	0.8487	0.9300	-	0.6943	-	0.9590
Att. WV + DGC	[31]	0.2220	-	0.1142	0.9300	-	0.5187	0.9619	0.9552
- + KNN	[31]	0.1944	-	0.6494	0.9100	-	0.6554	-	0.9318
- + KNN-A	[31]	0.2021	-	0.8595	0.9300	-	0.7549	-	0.9491
- + DW-KNN	[31]	0.2645	-	0.7083	0.9100	-	0.7248	-	0.9152
- + Cam-NN	[31]	0.1024	-	0.5145	0.9200	-	0.5364	-	0.9228
- + CNN	[31]	0.2571	-	0.7526	0.8900	-	0.7861	-	0.9491
SSMA + SFLDS	[31]	0.2731	-	0.7986	0.9300	-	0.6100	-	0.9145
PSO(4-2) + 5NN	[3]	-	-	-	-	-	-	0.9398	-
FFW-DGC	[33]	-	-	0.8958	0.9500	-	0.6252	0.8984	0.9745
MDis ABC	[15]	-	-	-	-	-	-		-
DCF + RFSC		0.4950	0.8552	0.8541	0.9500	0.6201	0.8101	0.9621	0.9916

TABLE V
COMPARATIVE MODEL SIZE ANALYSIS.

FS Method + Classifier	Ref.	Bupa	HillValley	Ionosphere	Iris	Musk1	Sonar	WDBC	Wine
FW FS + LDA	[39]	3.6	-	4.8	2.3	-	10.7	-	7.1
BW FS + LDA	[39]	4.7	-	30.4	3.9	-	56.4	-	12.8
PSO + LDA	[39]	4.6	-	21.7	3.6	-	38.1	-	12.3
PSO(4-2) + 5NN	[3]	-	12.22	3.26	-	76.54	11.24	3.46	6.84
MDis ABC	[15]	-	30.53	5.76	-	75.76	-	11.86	5.76
DCF + RFSC		5.8	8.3	16.4	3.1	46.2	25.8	11.5	3.3

information gathered from the extracted models, reinforcing the probability to extract the most promising regressors. Upon convergence a limit distribution is obtained which in practice identifies a single model structure. A distance correlation filtering (DCF) method has been occasionally found to be useful in reducing the feature set by pruning features that are independent from the model output.

The proposed method has been evaluated and compared to other well-known FS and classification algorithms obtaining quite promising and competitive results, especially in terms of the tradeoff between model complexity and classification accuracy. An important feature of the method is the easy interpretability of the obtained models, which can be used to gain more insight regarding the considered problem. Finally, the computational efficiency of the proposed method has been found to be comparable to that of competitor methods.

REFERENCES

[1] M. Dash and H. Liu, "Feature selection for classification," *Intelligent data analysis*, vol. 1, no. 1, pp. 131–156, 1997.

[2] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Computers & Electrical Engineering*, vol. 40, no. 1, pp. 16–28, 2014.

[3] B. Xue, M. Zhang, and W. N. Browne, "Particle swarm optimisation for feature selection in classification: Novel initialisation and updating mechanisms," *Applied Soft Computing*, vol. 18, pp. 261–276, 2014.

[4] H. Liu and H. Motoda, *Feature selection for knowledge discovery and data mining*, vol. 454. Springer Science & Business Media, 2012.

[5] F. Ferri, P. Pudil, M. Hatef, and J. Kittler, "Comparative study of techniques for large-scale feature selection," *Pattern Recognition in Practice IV*, pp. 403–413, 1994.

[6] M. Korenberg, S. Billings, Y. Liu, and P. McIlroy, "Orthogonal parameter estimation algorithm for non-linear stochastic systems," *International Journal of Control*, vol. 48, no. 1, pp. 193–210, 1988.

TABLE VI
COMPUTATION TIME $[s]$.

FS Method + Classifier	Ref.	Bupa	HillValley	Ionosphere	Iris	Musk1	Sonar	WDBC	Wine
PSO(4-2) + 5NN	[3]	-	1210.2	61.8	-	620.4	32.4	172.8	18.6
PSO + LDA	[39]	-	-	27.6	-	-	36.6	-	5.4
DCF + RFSC (Avrg.)		14.6	18.6	57	10.01	51.6	72	66	12

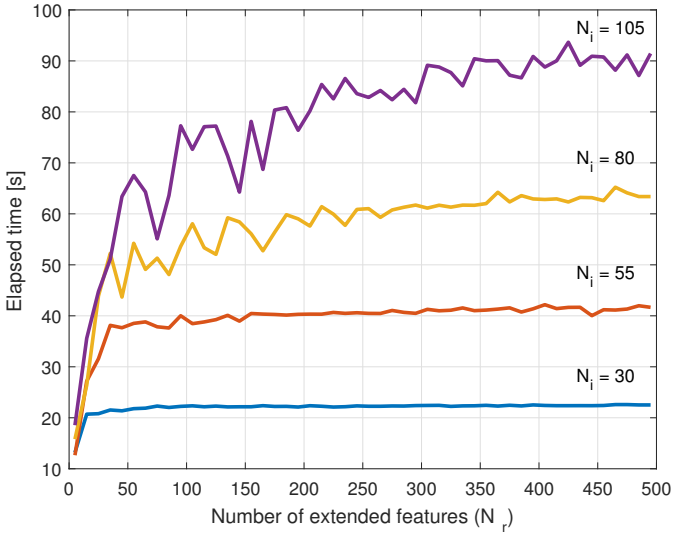

Fig. 4. Computational time of the RFSC algorithm for the WDBC dataset with $N_i = [30\ 55\ 80\ 105]$.

[7] L. Piroddi and W. Spinelli, "An identification algorithm for polynomial narx models based on simulation error minimization," *International Journal of Control*, vol. 76, no. 17, pp. 1767–1781, 2003.

[8] S. A. Billings, *Nonlinear System Identification: NARMAX Methods in the Time, Frequency, and Spatio-Temporal Domains*. Wiley, 2013.

[9] M. G. Smith and L. Bull, "Genetic programming with a genetic algorithm for feature construction and selection," *Genetic Programming and Evolvable Machines*, vol. 6, no. 3, pp. 265–281, 2005.

[10] J. Yang and V. Honavar, "Feature subset selection using a genetic algorithm," in *Feature extraction, construction and selection*, pp. 117–136, Springer, 1998.

[11] B. Xue, M. Zhang, and W. N. Browne, "Particle swarm optimization for feature selection in classification: A multi-objective approach," *IEEE Transactions on Cybernetics*, vol. 43, no. 6, pp. 1656–1671, 2013.

[12] M. M. Kabir, M. Shahjahan, and K. Murase, "A new hybrid ant colony optimization algorithm for feature selection," *Expert Systems with Applications*, vol. 39, no. 3, pp. 3747–3763, 2012.

[13] N. Sreeja and A. Sankar, "Pattern matching based classification using ant colony optimization based feature selection," *Applied Soft Computing*, vol. 31, pp. 91–102, 2015.

[14] R. Diao and Q. Shen, "Feature selection with harmony search," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 42, no. 6, pp. 1509–1523, 2012.

[15] E. Hancer, B. Xue, D. Karaboga, and M. Zhang, "A binary abc algorithm based on advanced similarity scheme for feature selection," *Applied Soft Computing*, vol. 36, pp. 334–348, 2015.

[16] S. Saha, S. Rajasekaran, and R. Ramprasad, "Novel randomized feature selection algorithms," *International Journal of Foundations of Computer Science*, vol. 26, no. 03, pp. 321–341, 2015.

[17] S. Mylavarapu and A. Kaban, "Random projections versus random selection of features for classification of high dimensional data," in *2013 13th UK Workshop on Computational Intelligence (UKCI)*, pp. 305–312, IEEE, 2013.

[18] M. Paliwal and U. A. Kumar, "Neural networks and statistical techniques: A review of applications," *Expert systems with applications*, vol. 36, no. 1, pp. 2–17, 2009.

[19] S. R. Gunn *et al.*, "Support vector machines for classification and regression," *ISIS technical report*, vol. 14, 1998.

[20] Y. Xu, "Maximum margin of twin spheres support vector machine for imbalanced data classification," 2016.

[21] D. W. Aha, D. Kibler, and M. K. Albert, "Instance-based learning algorithms," *Machine learning*, vol. 6, no. 1, pp. 37–66, 1991.

[22] P. G. Espejo, S. Ventura, and F. Herrera, "A survey on the application of genetic programming to classification," *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 40, no. 2, pp. 121–144, 2010.

[23] S.-W. Lin, K.-C. Ying, S.-C. Chen, and Z.-J. Lee, "Particle swarm optimization for parameter determination and feature selection of support vector machines," *Expert systems with applications*, vol. 35, no. 4, pp. 1817–1824, 2008.

[24] R. Sheikhpour, M. A. Sarram, and R. Sheikhpour, "Particle swarm optimization for bandwidth determination and feature selection of kernel density estimation based classifiers in diagnosis of breast cancer," *Applied Soft Computing*, vol. 40, pp. 113–131, 2016.

[25] I. Triguero, S. García, and F. Herrera, "Differential evolution for optimizing the positioning of prototypes in nearest neighbor classification," *Pattern Recognition*, vol. 44, no. 4, pp. 901–916, 2011.

[26] B. Li, Y. W. Chen, and Y. Q. Chen, "The nearest neighbor algorithm of local probability centers," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 38, no. 1, pp. 141–154, 2008.

[27] J. Wang, P. Neskovic, and L. N. Cooper, "Improving nearest neighbor rule with a simple adaptive distance measure," *Pattern Recognition Letters*, vol. 28, no. 2, pp. 207–213, 2007.

[28] S. A. Dudani, "The distance-weighted k-nearest-neighbor rule," *IEEE Transactions on Systems, Man and Cybernetics*, no. 4, pp. 325–327, 1976.

[29] Q.-B. Gao and Z.-Z. Wang, "Center-based nearest neighbor classifier," *Pattern Recognition*, vol. 40, no. 1, pp. 346–349, 2007.

[30] C. Y. Zhou and Y. Q. Chen, "Improving nearest neighbor classification with cam weighted distance," *Pattern Recognition*, vol. 39, no. 4, pp. 635–645, 2006.

[31] A. Cano, A. Zafra, and S. Ventura, "Weighted data gravitation classification for standard and imbalanced data," *IEEE Transactions on Cybernetics*, vol. 43, no. 6, pp. 1672–1687, 2013.

[32] L. Peng, B. Yang, Y. Chen, and A. Abraham, "Data gravitation based classification," *Information Sciences*, vol. 179, no. 6, pp. 809–819, 2009.

[33] L. Peng, H. Zhang, H. Zhang, and B. Yang, "A fast feature weighting algorithm of data gravitation classification," *Information Sciences*, 2016.

[34] A. Falsone, L. Piroddi, and M. Prandini, "A randomized algorithm for nonlinear model structure selection," *Automatica*, vol. 60, pp. 227–238, 2015.

[35] G. J. Székely, M. L. Rizzo, and N. K. Bakirov, "Measuring and testing dependence by correlation of distances," *Ann. Statist.*, vol. 35, pp. 2769–2794, 12 2007.

[36] C. Bishop, *Pattern Recognition and Machine Learning*. Springer, New York, 2006.

[37] D. Newman, S. Hettich, C. L. Blake, and C. J. Merz, "UCI repository of machine learning databases," 1998.

[38] A. Ben-David, "Comparison of classification accuracy using Cohen's weighted kappa," *Expert Systems with Applications*, vol. 34, no. 2, pp. 825–832, 2008.

[39] S.-W. Lin and S.-C. Chen, "PSOLDA: A particle swarm optimization approach for enhancing classification accuracy rate of linear discriminant analysis," *Applied Soft Computing*, vol. 9, no. 3, pp. 1008–1015, 2009.

Aida Brankovic was born in Sarajevo (Bosnia and Herzegovina), in 1987. In 2009 she received her bachelor degree and in 2011 her master degree, both in Automation, Control and Electronics from University of Sarajevo. From February to September 2013 she worked as a teaching assistant at the Electrical Faculty of University of Sarajevo, and from September to November 2013 as research assistant in the MOVE research group of the Politecnico di Milano. In November 2013 she started her PhD at the Dipartimento di Elettronica, Informazione e Bioingegneria of the Politecnico di Milano, Systems and Control section. Her current interests include nonlinear model identification, randomized algorithms and supervised machine learning.

Alessandro Falsone received the Bachelor of Science in 2011 and the Master of Science cum laude in 2013, both in Automation and Control Engineering from Politecnico di Milano. From November 2013 to October 2014 he worked as a research assistant at the Dipartimento di Elettronica, Informazione e Bioingegneria at Politecnico di Milano. Since November 2014 he is a Ph.D. student in the System and Control division of the same department. His current research interests include distributed optimization and control, optimal control of stochastic hybrid systems, randomized algorithms, and nonlinear model identification.

Maria Prandini received her laurea degree in Electrical Engineering (summa cum laude) from Politecnico di Milano (1994) and her Ph.D. degree in Information Technology from Universitá degli Studi di Brescia, Italy (1998). From 1998 to 2000 she was a postdoctoral researcher at the Dep. of Electrical Engineering and Computer Sciences, Univ. California at Berkeley. She also held visiting positions at Delft Univ. of Technology (1998), Cambridge Univ. (2000), Univ. of California at Berkeley (2005), and Swiss Federal Inst. of Technology Zurich (2006). In 2002, she started as an Assistant Professor in Systems and Control at Politecnico di Milano, where she is currently an Associate Professor. Her research interests include stochastic hybrid systems, randomized algorithms, constrained control, system abstraction and verification, nonlinear identification, distributed optimization, and the application of control theory to air traffic management and energy systems. She serves on the editorial board of Cyber Physical Systems, and previously of European Journal of Control, IEEE Trans. on Automatic Control, IEEE Trans. on Control Systems Technology and Nonlinear Analysis: Hybrid Systems. From 2013 to 2105, she has been editor for Electronic Publications of the IEEE CSS. She is member of the IEEE CSS Board of Governors, and since January 2016 she is CSS Vice-President for Conference Activities.

Luigi Piroddi (M'07) was born in London, U.K., in 1966. He received his laurea degree in Electrical Engineering and the Ph.D. degree in Computer Science and Control Theory from the Politecnico di Milano, Milano, Italy, in 1990 and 1995, respectively. Between 1994 and 1999, he was a Professor of fundamentals of automation with the Universitá degli Studi di Bergamo, Bergamo, Italy. From 1999 to 2004, he was an Assistant Professor with the Politecnico di Milano. From 2004 to 2015 he has been an Associate Professor, and from 2016 he is Full Professor with the same institution, where he holds various courses in the systems and control area. His research interests include nonlinear model identification, Petri nets, modeling, and control of manufacturing processes. He currently serves on the editorial board of the IEEE Transactions on Automation Science and Engineering.