Enhancing Genetic Feature Selection Through Restricted Search and Walsh Analysis

Sancho Salcedo-Sanz, *Member, IEEE*, Gustavo Camps-Valls, Fernando Pérez-Cruz, *Member, IEEE*, José Sepúlveda-Sanchis, and Carlos Bousoño-Calzón, *Member, IEEE*

***Abstract*—In this paper, a twofold approach to improve the performance of genetic algorithms (GAs) in the feature selection problem (FSP) is presented. First, a novel genetic operator is introduced to solve the FSP. This operator fixes in each iteration the number of features to be selected among the available ones and consequently reduces the size of the search space. This approach yields two main advantages: a) training the learning machine becomes faster and b) a higher performance is achieved by using the selected subset. Second, we propose using the Walsh expansion of the FSP fitness function in order to perform ranking on the problem features. Ranking features have been traditionally considered to be a challenging problem, especially significant in health sciences where the number of available and potentially noisy signals is high. Three real biological datasets are used to test the behavior of the two approaches proposed.**

***Index Terms*—Diabetes mellitus, feature selection, filter methods, genetic algorithms, thrombin binding, unstable angina, wrapper methods.**

I. Introduction

The FEATURE selection problem (FSP) is an open issue in machine learning, which basically consists of finding a subset of input features that describes the underlying system structure as well or better than all available features. There is a more general problem, known as feature extraction problem (FEP) [1], in which one has to construct a few features that encode all of the information contained in the available original ones; the FSP is, in fact, a particularization of this problem, which we focus on in this paper.

In fact, before building a model, the most significant predictors must be selected; otherwise, insignificant features could become noise and alter its performance, thus producing an unreasonable outcome. This is especially true when the number of available input variables is large, and exhaustive search through all combinations of features is computationally infeasible. This fact is intimately related to the *curse of dimensionality*, and its Hughes attendant [2], given that a huge number of features induces a computational expensive learning process [3] and, in turn, this produces suboptimal models due to the quality of data used. These are certainly main issues to be addressed and a feature selection stage is a common choice.

The relevance of the FSP appears when the given features are used to explain the achieved results. This relates to that, in some applications such as bioinformatics or health sciences, being able to explain the obtained solution (in terms of the selected input features) becomes as relevant as obtaining the best possible answer (accuracy of the subsequent classifier or regressor). In medical or bioinformatics applications, the feature selection is also relevant because data are usually very scarce compared to the number of features (even order of magnitudes) and, therefore, overfitting will probably occur, significantly reducing the performance of the system [4], [5].

The search for the best m features out of the n available is known to be an NP-complete problem (it cannot be solved in polynomial time unless that $\mathrm{P} = \mathrm{NP}$) [6] and the number of local minima can be quite large. Many methods have been used for solving the FSP: pruning methods for neural network [7]–[10]; mutual information techniques [11], [12]; incremental learning [13]; principal/independent component analysis [14], [15] classification trees [16]; self-organizing maps [17]; fuzzy clustering [18]; etc. Nevertheless, following a general taxonomy, feature selection can be divided into two major categories: filter methods [19] and wrapper [4] methods. The former use an indirect measure of the quality of the selected features, so a faster convergence of the algorithm is obtained. On the other hand, wrapper methods use as selection criteria the output of the learning machine. This approach guarantees that in each step of the algorithm, the selected subset improves the performance of the previous one. Filter methods might fail to select the right subset of features if the used criterium deviates from the one used for training the learning machine, whereas wrapper methods can be computationally intensive since the learning machine has to be retrained for each new set of features.

In this paper, we use genetic algorithms (GAs) [20] in order to solve the FSP. GAs, under certain conditions, are able to find the global optimum of a multiple local-minima problem, and have already been used to solve the FSP as a filter [21], [22] and a wrapper [4] method, obtaining good results in the number of selected features and in the overall performance of the classifier. However, for large dimensional problems, as the ones encountered in bioinformatics, the GA approach has failed to converge

Manuscript received January 31, 2003; revised April 14, 2004. This work was supported by the CICYT under Project Number TIC2000-0380-C03-03. A reduced version of this paper was presented in the workshop *Learning '02*, Leganés, Madrid, 2002. This paper was recommended by Guest Editors J. L. Fernández and A. Artés.

S. Salcedo-Sanz, F. Pérez-Cruz, and C. Bousoño-Calzón are with the Department of Signal Theory and Communications, Universidad Carlos III de Madrid, Leganés-Madrid 28911, Spain (e-mail: sancho@tsc.uc3m.es; fperez@tsc.uc3m.es; cbousono@tsc.uc3m.es).

G. Camps-Valls is with Grup de Processament Digital de Senyals, Department of Enginyeria Electrònica, Universitat de València, València 46100, Spain (e-mail: gustavo.camps@uv.es).

J. Sepúlveda-Sanchis was with the Grup de Processament Digital de Senyals, Dept. Enginyeria Electrónica, Universitat de València (Spain). He is now with the Center for Genomics and Bioinformatics (CGB), Stockholm, Sweden.

Digital Object Identifier 10.1109/TSMCC.2004.833301

or, if it does, its performance has been poor when compared to other methods. In order to alleviate these problems, we present a twofold strategy to improve performance of GAs in the FSP:

- *The m-features operator*. First, we present a novel operator that it is able to reduce the size of the search space. This operator will force the number of selected features to remain constant in each iteration, which allows to select a specific number of features beforehand. This has several additional advantages: first the algorithm will converge faster, and second, the operator can be used both over filter or wrapper methods.
- *Feature ranking using the Walsh expansion of the GA fitness function*. Second, we propose the use of a modified spectrum [23] of the FSP fitness function, which is derived from its Walsh expansion [24]. This measure will allow us to perform a ranking of the available features and, consequently, to gain knowledge in the problem by identifying relevant or meaningless features.

We have applied these procedures in three relevant medical and bioinformatics problems. The first one is the so-called *thrombin binding problem* for drug discovery, which was used in the *Knowledge Discovery and Data Mining* (KDD) Cup 2001. The second problem is related to a relevant issue in coronary diseases: the prognosis of mortality risk in patients who suffer from unstable angina. The third problem deals with the discrimination of patients with signs of diabetes according to World Health Organization criteria.

The rest of the paper is organized as follows. We describe the FSP in detail in Section II. In Section III, we review the GAs and the Walsh Analysis to study the spectrum of a GA fitness function. The operator used to fix the number of selected features is detailed in Section IV, along with the *modified spectrum* for ranking features. Section V is devoted to show, by means of computational experiments over biological data, the results achieved with the proposed modifications. We close the paper in Section VI by providing some concluding remarks.

II. Feature Selection Problem

The FSP in a learning from samples scheme can be addressed as follows. Given a set of labeled data points $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_l, y_l)\}$, where $\mathbf{x}_i \in \mathbb{R}^n$ and $y_i \in \{\pm 1\}$, choose a subset of m features $(m < n)$, that achieves the lowest classification error. Following [6], we will define the FSP as finding the optimum n-column vector $\boldsymbol{\sigma}$, where $\sigma_i \in \{1, 0\}$, that defines the subset of selected features, is found as

$$\boldsymbol{\sigma}^o = arg \min_{\boldsymbol{\sigma}, \boldsymbol{\alpha}} \left(\int V\left(y, f(\mathbf{x} * \boldsymbol{\sigma}, \boldsymbol{\alpha})\right) dP(\mathbf{x}, y) \right) \qquad (1)$$

where $V(\cdot, \cdot)$ is a loss functional, $P(\mathbf{x}, y)$ is the unknown probability function the data were sampled from and we have defined $\mathbf{x} * \boldsymbol{\sigma} = (x_1\sigma_1, \ldots, x_n\sigma_n)$. The function $y = f(\mathbf{x}, \boldsymbol{\alpha})$ is the classification engine that is evaluated for each subset selection $\boldsymbol{\sigma}$ and for each set of its hyper-parameters $\boldsymbol{\alpha}$.

In such an approach, the objective is to process the data in order to extract valid, novel, potentially useful, and an ultimately understandable structure in data by identifying relevant and meaningless features [25]. This is the first step in a *knowledge discovery* learning scheme. In this context, two main approaches can be followed:

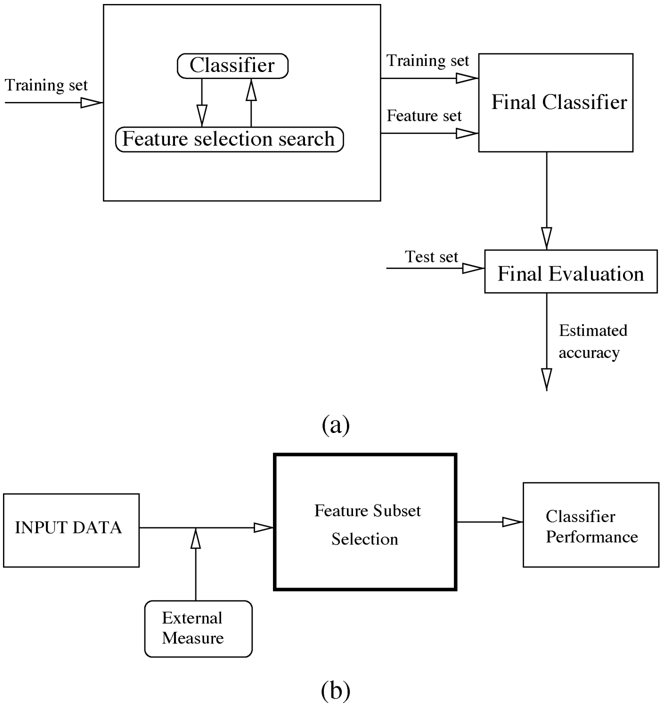

Fig. 1. (a) Outline of a wrapper method. (b) Outline of a filter method.

- The *wrapper approach* to the FSP was introduced in [26]. In this approach, the feature selection algorithm conducts a search for a good subset of features using the classifier itself as part of the evaluating function. Fig. 1(a) shows the idea behind the wrapper approach: the classifier is run on the training dataset with different subsets of features. The feature subset, which produces the lowest estimated error in an independent but representative test set, is chosen as the final feature set. For further considerations about wrappers methods, the following bibliography can be consulted [4], [5], [27].
- In the *filter approach* to the FSP, the feature selection is performed based on the data, ignoring the classifier algorithm. An external measure calculated from the data must be defined in order to select a subset of features. After the search, the best feature subset found is evaluated on the data by means of the classifier algorithm. Note that the filter algorithms performance completely depends on the measure selected for comparing feature subsets. Fig. 1(b) shows an example of how a filter algorithm works.

  Filter methods are usually faster than wrapper methods. However, their main drawback is that they totally ignore the effect of the selected feature subset on the performance of the classification algorithm during the search. Further analysis and application of filter methods can be found in [21] and [22].

For both wrapper and filter methods, a binary representation can be used for the FSP, where a 1 in the ith position of the binary vector means that the feature i is considered within the subset of features, and a 0 in the jth position of the binary vector means that feature j is not considered within the subset of features. Note that using this notation is equivalent to encode the problem as the vector $\boldsymbol{\sigma}$ included in expression (1). Note also that there are 2^n different subsets of features (being n the total number of features), and the problem is to select the best one in terms of a certain measure, which can be either internal (wrapper methods) or external (filter methods) to the classifier.

III. Genetic Algorithms

GAs [20] are a class of robust problem-solving techniques based on a population of solutions, which evolve through successive generations by means of the application of three genetic operators: selection, crossover, and mutation [20]. GAs are suited to perform a search in huge search spaces, where other methods (local or gradient searches) cannot provide good results. The FSP encoded with a binary representation is one of these cases.

A. *Walsh Analysis and Spectrum*

Walsh Analysis of a function is a commonly used method in GAs to study the internal structure of the fitness functions [24]. It has also been used to explain how a GA works. Walsh Analysis is equivalent to a Fourier expansion of a function in the binary search space $\{0,1\}^n$. The Walsh expansion of a function associates a Walsh coefficient w_j to a binary vector $\mathbf{j}$ (*partition*). The function can be completely reconstructed from partitions[1] $\mathbf{j}$s and Walsh coefficients w_j. Continuing with the notation, we give the main steps to define Walsh expansion of a function.

The Walsh basis function for a partition $\mathbf{j}$, $\psi_{\mathbf{j}} : \{0,1\}^n \to \mathbb{R}$ is defined as

$$\psi_{\mathbf{j}}(\mathbf{x}) = \prod_{i=1}^{n} (-1)^{x_i j_i} \tag{2}$$

where x_i and j_i are the components of binary vectors $\mathbf{x}$ and $\mathbf{j}$.

Walsh functions form a complete orthogonal set of basis functions [20]. Every function $f : \{0,1\}^n \to \mathbb{R}$ can be expanded as

$$f(\mathbf{x}) = \sum_{j=0}^{2^n-1} w_j \psi_{\mathbf{j}}(\mathbf{x}) \tag{3}$$

where

$$w_j = \frac{1}{2^n} \sum_{x=0}^{2^n-1} f(\mathbf{x})\psi_{\mathbf{j}}(\mathbf{x}). \tag{4}$$

The Walsh expansion captures the internal structure of a function: if this function has dependencies among variables, then its Walsh coefficients for partitions involving nondependent variables are zero. For example, let $f(x_1,x_2,x_3,x_4) = f_1(x_1,x_2) + f_2(x_3,x_4)$, then $w_{0111} = w_{1110} = w_{0110} = 0$ [28]. For further analysis and details on Walsh Analysis, see [24].

The *spectrum* of a function [23] is a graphic representation of the most important partitions of the function which is obtained from its Walsh expansion. The *order* of a partition $\mathbf{j}$ is defined as the number of 1s in it. Note that in the FSP with binary representation, the order of a partition is equivalent to the number of selected features.

Using the definitions above, the spectrum of a function is defined starting from its Walsh coefficients as follows. Let $\wp$ be the set of all partitions belonging to the search space $S = \{0,1\}^n$. Let $\wp_p$ be the set of partitions belonging to S with order p. A total energy for the function is defined as

$$\sigma^2 = \sum_{j \in \wp} w_j^2. \tag{5}$$

The energy for the partitions with order p is

$$\beta_p^2 = \sum_{j \in \wp_p} w_j^2 \tag{6}$$

and their normalized energy

$$B_p = \frac{\beta_p^2}{\sigma^2}. \tag{7}$$

Vector $\mathbf{B} = \{B_1 \ldots B_n\}$ is the *spectrum* associated to the Walsh expansion of $f(\cdot)$. It can be readily shown that $B_p \geq 0$ and $\sum_p B_p = 1$. In this paper, we will use a modification of the spectrum in order to perform a ranking of the best features for a given FSP problem.

IV. Enhancing Genetic Feature Selection

GAs have been proposed and successfully used for the FSP, both as wrapper and filter methods [4], [5], [21], [22], [27], but its use has been limited mainly due to convergence problems and poor performance in large features spaces. Nevertheless, few efforts have been done to investigate on how to improve the performance of a GA in FSP. From our particular point of view, a better adaptation of the GA to the FSP is therefore necessary. In this section, we present two techniques in order to introduce a major adaptation of the GA to the FSP. First, we propose reducing the search space by means of the *m-features* operator. Second, we present a modified spectrum for ranking features.

A. *m-Features Operator*

One of the main difficulties that a GA faces when tackling FSP in high dimensional input spaces is that convergence can be slow and additionally its performance becomes poorer than expected. This problem is even bigger in wrapper methods, where the calculation of the fitness function is provided directly by the classifier, what introduces a considerable time delay in every generation of the GA. To overcome this problem, we propose a novel genetic operator which fixes the number of features selected by the GA to m features (i.e., the number of 1s in the individuals remains constant to a given number m).

The so-called *m-features* operator works in the following way: after the application of the crossover and mutation operators, a given individual $\boldsymbol{\sigma}_i$ of the GA population will present p 1s that, in general, will be different from the desired number of features m. If $p < m$, the *m-features* operator adds $(m-p)$ 1s randomly and if $p > m$, the *m-features* operator randomly selects $(p-m)$ 1s and removes them from the binary string. The *m-features operator* can be described in pseudocode, as follows.

[1] Hereafter, we will denote, in boldface, a partition indexed as a binary vector $\mathbf{j}$ and in normal type j, its corresponding integer value.

```
The $m$-features operator
Select $m$ (number of features) before
running the GA.
for every generation of the GA:
  for every individual $\boldsymbol{\sigma}_i$ of the GA
population:
    check the number of 1s $p$.
      if ($p < m$)
        Add_ones($m - p$);
      else
        Remove_ones($p - m$)
      end(if)
  end(individual $\boldsymbol{\sigma}_i$)
end(generation)
```

Note that the m*-features* operator forces the GA to search in a reduced search space. In fact, the standard GA searches for vector $\boldsymbol{\sigma}^\circ$ in a space of size 2^n, whereas introducing the m*-features* operator, the size of the search space is reduced to $\binom{n}{m} \ll 2^n$. In addition, the application of the m*-features* operator will consequently obtain a much faster convergence of the GA.

Three specific actuation modes can be obtained for this new genetic operator. The GA can be used with the m*-features* operator (*best* m*-features* mode), using the $(m_1 - m_2)$-features operator (being the number of features lower bounded by m_2 and upper bounded by $m_1 > m_2$), or without any restricted search (the GA is running in *standard* mode, $m_1 = 0$ and $m_2 = n$). The *standard* mode will search for the best set of features in the whole search space, whereas the GA running in *best* m*-features* mode searches for the best set of m features in the search space. Note that the m*-features* operator is an adaptation of the GA to the FSP, due to the number of 1s in the individuals is the number of selected features. Note also that the m*-features* operator can be used both in wrapper and filter methods, as will be shown in the experiments section.

B. Walsh Analysis for Ranking of Features

The spectrum of a function defined in [23] (also summarized in Section III-A) represents the distribution of energies among different orders (number of 1's) of the partitions which form the search space. If the analyzed function is the fitness function of an FSP, the spectrum might give a measure of what are the most important features (i.e., it could be used as a tool to perform a ranking of features in a given FSP).

The definition of spectrum from the Walsh Analysis of a function given in [23] calculates the energy associated to partitions of the same order. For example, in the search space $\{0,1\}^4$, partitions $\{1011\}$, $\{1110\}$, $\{0111\}$ contribute to the same component of the spectrum (B_3); however, they represent three different sets of features. Thus, the "classical" definition of spectrum cannot be used for ranking the features according to their relevance. In order to solve this problem, we propose a slight modification of the definition of the spectrum, in the following way.

Let ζ_i be the set of partitions $\mathbf{j}$ with a 1 in the position i_{th}. In the example above partitions $\{1011\}$ and $\{1110\}$ belong to ζ_1 (they also belong to ζ_3) whereas partition $\{0111\}$ does not belong to ζ_1 (but it belongs to ζ_3). Therefore, a modified spectrum, named the *prime spectrum*, can be defined as follows:

$$B'_i = \frac{\sum\limits_{j \in \zeta_i} w_j^2}{\sum\limits_{\forall j} o(\mathbf{j}) \cdot w_j^2} \tag{8}$$

where $o(\mathbf{j})$ is the order of $\mathbf{j}$.

Prime spectrum fulfills $B'_i \geq 0$ and $\sum_i B'_i = 1$. In fact, it can be interpreted as the associated energy to every feature in the binary search space and, thus, features with large values of B'_i are more relevant than features with small values of B'_i. Consequently, we propose using vector $\mathbf{B}'$ to perform ranking of features in the FSP.

Note that vector $\mathbf{B}'$ depends on the fitness function selected for the FSP through the values of its Walsh coefficients. Note that vector $\mathbf{B}'$ can be calculated both for wrapper and filter methods. However, in large search spaces, the calculation of the Walsh expansion can be computationally infeasible, and estimation methods as the one proposed in [23] should be used.

V. Experiments

Performance of our proposals is evaluated in three real datasets. The first one is a real application in molecular bioactivity for drug design, also known as thrombin binding problem (TBP) used for the first time in the KDD cup in 2001. In this application, we test the performance of the m*-features* operator in a GA used as a wrapper feature selection method. The second experiment deals with the assessment of six months ahead mortality risk due to *angina pectoris*. In this application, we test the performance of the m*-features* operator in a GA used as filter method. We also test here the ranking of features obtained through the prime spectrum, and compare it with commonly used sensitivity measures on the best classifiers. The third problem deals with the discrimination of patients with signs of *diabetes mellitus*. We compare here the performance of a multilayer perceptron (MLP) with and without a feature selection stage based on the prime spectrum and sensitivity measures. For all experiments, we have used a standard GA as defined in Section III, with a population size $\xi = 25$ individuals, and probabilities of crossover and mutation $P_c = 0.6$ and $P_m = 0.01$, respectively.

A. Molecular Bioactivity for Drug Design

The first step in the discovery of a new drug is to identify and isolate the receptor to which it should bind, followed by testing many molecules for their ability to bind to the target site. Thus, it needs to be determined as to what features of the drug molecule separate the active (binding) compounds from the inactive (nonbinding) ones. In this case, the problem consists of selecting the features of a drug molecule that make it bind to a target site on thrombin, a key receptor in blood. This task defines the TBP, which can be seen in a machine learning context as a FSP.

1) Dataset: The dataset used in this paper is a TBP provided by *DuPont Pharmaceutical* for the KDD cup 2001 Competition

[29]. Each example (observation) has a length of 139 351 binary features which describe three-dimensional (3-D) properties of the drug molecule. Note that this is the only information given by the dataset provider concerning the nature of the features. The reason for this was that the competition would evaluate the effectiveness of the algorithms without the advantage of being gained from the designer's prior biological knowledge. Following the winner criteria explained in [29], we selected 100 features from the original set of 139 351 using mutual information. Every example (observation) is labeled as positive $+1$ of negative -1, depending on whether the example binds or not, respectively.

The dataset was then split into a training and a test set. In the training set, there are 1909 examples (100 features) and only 42 of them bind. Hence, the data are highly unbalanced (42 positive examples is only 2.2% of the data). The test set contains 634 additional compounds (also 100 features), which were generated based on the assay results recorded for the training set.

2) Genetic Feature Selection for the TBP: In this problem, we have applied the GA as a wrapper method for FSP. Note that the size of the searching space is 2^{100}, so we use the *m-features* operator in order to reduce it and obtain a faster convergence of the GA. Due to the characteristics of the problem's features as binary numbers, we selected one of the simplest classifiers as inductive classification algorithm: the OR classifier. This classifier is defined in the following way:

$$f(\mathbf{x}) = \begin{cases} 1, & \text{if } \frac{\sum_{i=1}^{n} x(i)}{n} > 0 \\ -1, & \text{otherwise} \end{cases} \tag{9}$$

where $f(\mathbf{x})$ assigns the prediction that a new molecule binds if any of the selected features (represented by vector $\mathbf{x}$, $x(i)$ the i_{th} feature) were nonzero, and n is the number of features.

The use of this operator is fully justified because there are very few 1s in each feature and we are looking for features with 1s highly correlated with the positive outputs. In addition, the OR classifier provides a very fast way of computing the percentage of error in classification for this problem, without training or adjusting any hyper-parameter.

Following the KDD rules, the error in classification is evaluated according to a *weighted accuracy criterion*, due to the unbalanced nature of the number of positive and negative examples

$$\begin{aligned} err_{wac} = \frac{1}{2}\left(\frac{\#\{\hat{y} : y = 1 \wedge \hat{y} = 1\}}{\#\{y : y = 1\}}\right) \\ + \frac{1}{2}\left(\frac{\#\{\hat{y} : y = -1 \wedge \hat{y} = -1\}}{\#\{y : y = -1\}}\right) \end{aligned} \tag{10}$$

where y stands for the correct label of a sample, and $\hat{y}$ stands for the output of the classifier for that sample. Note that complete classification success is scored with $err_{wac} = 1$.

This classification error is used to calculate the GA fitness function. Since the GA works by maximizing the fitness, and the problem objective involves minimizing err_{wac}, we have defined the GA fitness function in the following way:

$$fitness(\boldsymbol{\sigma}_i) = 100(1 - err_{wac}). \tag{11}$$

TABLE I
BEST err_{wac} OBTAINED BY GA RUNNING IN DIFFERENT MODES AND THE BEST err_{wac} IN THE KDD CUP 2001

Algorithm	err_{wac}
GA (*7-features* mode)	70.8%
GA (*10-features* mode)	71.1%
GA (*13-features* mode)	**74.5%**
GA (*16-features* mode)	73.5%
GA (*Standard* mode)	69.6%
Winner of KDD Cup 2001 [29]	68.4%

where $\boldsymbol{\sigma}_i$ represents every individual in the GA population. Note that err_{wac} depends on the individual $\boldsymbol{\sigma}_i$.

3) Obtained Results: We compare the results obtained by the GA as a wrapper method with the results provided by the KDD cup 2001 competition [29]. Only 7% of all the competitors at KDD achieved an err_{wac} higher than 60.0%. The winner used a *Bayes Network*; to be more precise, tree augmented Naive-Bayes networks. In [30] and [31], the authors describe in more detail the used techniques for the Bayes networks.

Table I shows the results of the best solution achieved by our GA running in *best 7-features*, *10-features*, *13-features*, and *16-features* modes, *standard* mode, and the results obtained for the winner of the KDD.

Note that our algorithm achieves better results that the best existing algorithm. The best solution was obtained with the GA working in *best 13-features* mode, with a $err_{wac} = 74.5\%$, an improvement of 6% over the best solution given in the KDD Cup. The best solution obtained with the GA working in *standard* mode selected 18 features, with a $err_{wac} = 69.6\%$, which is still better than the best solution in the KDD Cup. However, in general, the GA running in *standard* mode performs worse than running in *m-feature* mode.

B. Risk Assessment of Unstable Angina

Angina is the primary symptom of coronary artery disease and, in severe cases, of a heart attack. Angina is usually referred to as stable (predictable) or unstable (less predictable and a sign of a more serious situation). In this context, the use of classification methods helps to predict mortality due to angina.

1) Data Collection: The Recursos Empleados en el Síndrome Coronario Agudo y Tiempos de Espera (RESCATE) study consisted of a registry of first acute myocardial infarct (AMI) and unstable angina (UA) patients admitted to one hospital with, and three others without, coronary angiography facilities or coronary surgery.[2] A total of 2661 patients with unstable angina were consecutively admitted to the participating hospitals.

2) Learning Scheme and Results: The learning scheme followed in this problem is basically constituted by three stages.

- *Stage I: Feature selection*. In this application, we have used several information criteria as fitness functions for the GA: the Mallow's C_p criterion, the classical Akaike's information criteria (AIC) and the maximum description length (MDL) criteria. As in the previous application,

[2] More information regarding this study can be retrieved from http://www.imim.es.

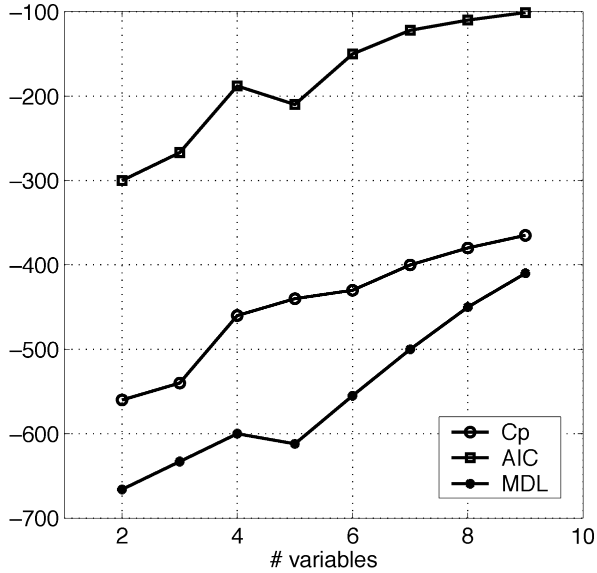

Fig. 2. Evolution of Mallow's C_p, AIC and MDL information criteria for different subsets of variables.

we have transformed these the-smaller-the-better performance statistic measurements into fitness functions to be maximized by previously rescaling them. This approach is based on the work [21], where the fitness function is chosen to be the Mallow's C_p statistical criterion, as applied to the binary coded linear regression $\mathbf{y}_i = a\mathbf{x}_i + b$ where $a, b \in \{0, 1\}$. In this case, each possible subset can be described as a binary string of length n. However, we introduce the important modification of fixing the number of features in the solution coding at each iteration by means of the *m-features operator*.

First, we selected relevant variables through conventional feature selection techniques (principal component analysis, correlation function, statistical descriptors, and entropy measures). The best subset was reduced to 14 variables from the originally 75 collected. However, we continued inspecting smaller subsets with the GA approach (Fig. 2). Despite the optimal subset being constituted by only two variables for all measures (relatives with schemic cardiopathy and previous diagnosis of angina), there are no significant differences between the solution of five factors (relatives with schemic cardiopathy, previous diagnosis of angina, hyper-cholesterolemia, habitual smoker, and gender). The latter selection seems to be more robust clinically [32] and, additionally, AIC and MDL criteria show a local minimum which suggests that selection as well. However, this must be verified by developing dedicated models for each possible subset and evaluating their performance in a test set.

- *Stage II: Classification*. We have evaluated the quality of the selection by developing dedicated SVM [33] for several subsets identified in the previous stage. This step has two particular advantages; first, it allows to assess previous selection methods and second, it provides an accurate and robust classifier. Data were split into a training set (483; 22 cases with UA) and a validation set (243 patients; 12 of cases with UA). Selection of the model was subjected to tight restrictions: first, the negative predictive value (NPV%) should be higher than 97.5% since the rate of false predictions on the true positives should be avoided as much as possible; and second, well-balanced models are desirable so we used the err_{wac} measure [see (10) for details].

TABLE II
RESULTS IN THE VALIDATION SET OF SVMS WITH DIFFERENT SUBSETS. THE VALUE N INDICATES THE NUMBER OF VARIABLES IN EACH SUBSET. IN ALL CASES, THE BEST SVM CONSIDERED AN RBF KERNEL

Score	**N=2**	**N=5**	**N=14**	**N=75**
err_{wac}	72.62%	74.84%	72.11%	66.45%
NPV	97.87%	98.56%	97.87%	88.57%

Table II shows results obtained for the selected subsets of variables, where N indicates the number of variables in each subset. The best SVM model (RBF kernel, $N = 5$) yielded good results ($err_{wac} = 74.84\%$ in the validation set) although regularization was a hard problem to solve and, thus, a different individualization parameter was used for each class, as proposed in [34]. In addition, a high rate of negative predictive values $(\mathrm{NPV} = 98.56\%)$ was obtained.

- *Stage III: Ranking*. Once the most relevant feature subset has been selected with the previous methods, a Walsh expansion of the GA fitness function is performed and the corresponding prime spectrum is calculated. This provides a ranking of variables which is compared to that from a sensitivity analysis of the best classifier. Sensitivity analysis (SA) is commonly used to study the influence of input variables in a classifier. Candidate models are constructed by evaluating the effect of removing an input variable. This measure, commonly known as *delta error* (DE) in the literature, produces a valuable ranking of variables relevance. Two additional sensitivity measures can be computed based on perturbing an input and monitoring model output observations: the *average gradient (AG)* and the *average absolute gradient (AAG)* [35].

At this stage, we inspect the ranking provided by the prime spectrum obtained from the Walsh expansion of GA fitness functions, using the final 14 variables [Fig. 3(a)]. An additional measure of the feature relevance is to consider all partitions which order comprises more than 90% of the spectrum energy. In our case study, an order of $m = 9$ is found. This indicates that partitions with higher order contributes with negligible energy (i.e., no more than nine variables are necessary to describe the problem accurately). Moreover, no significant difference is found between $m = 9$ and $m = 5$ regarding the 0.9-quantile of the FSP (0.935 and 0.899, respectively). These results match the ones obtained with statistical criteria in stage I. In Fig. 3(b), we show ranking of variables according to three common sensitivity measures for the best SVM with 14 input variables. Results perfectly match the ones obtained with the Walsh expansion, which indicates that a robust classification model has been achieved and confirms the final selection of five features.

These are promising results from a clinical viewpoint. Variables selected have been reported in the literature to

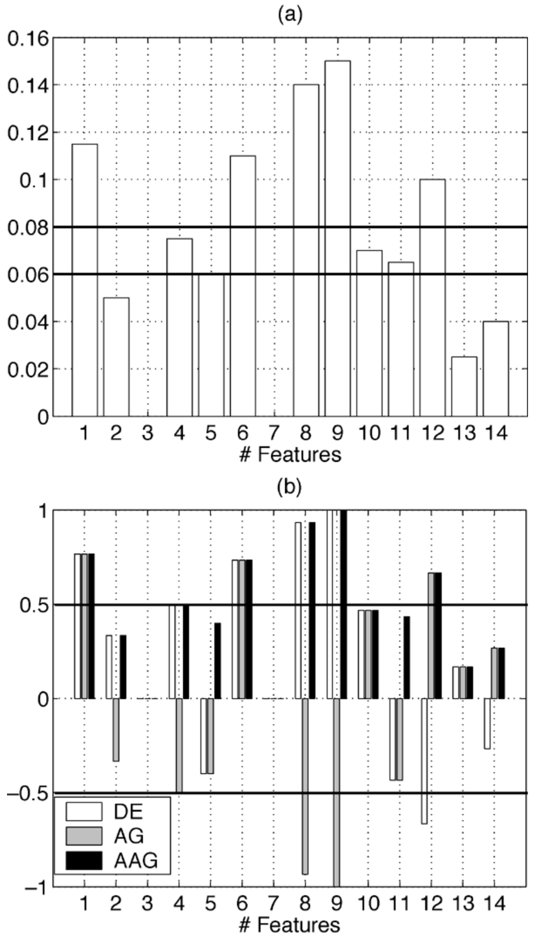

Fig. 3. (a) Ranking features attending to the prime spectrum $\mathbf{B}'$. The horizontal lines indicate two different thresholds for subset selection. (b) Ranking provided by sensitivity measures of the best SVM with $N = 14$. The ±50% relevance thresholds are indicated with horizontal lines.

be of fundamental relevance for predicting nonfatal acute myocardial infarction or death six months later in unstable angina *pectoris* [32]. However, more efforts must be done to define efficiently the energy distribution drawn from the Walsh expansion.

This methodology permits to assess GA-driven feature selection results and allows to obtain accurate and robust classification machines. Additionally, some knowledge gain in the problem and important clinical conclusions are derived.

C. Diabetes Detection

The diagnosis of signs of diabetes according to World Health Organization criteria (i.e., if the 2-h postload plasma glucose was at least 200 mg/dl at any survey examination or if found during routine medical care) is considered as the third real example. Data were downloaded from the UCI repository corresponding to the pima-indians-diabetes dataset. We used a similar approach as in the previous example but an MLP was used as classifier in Stage II.

Two feature selection methods were considered; the one drawn by the Walsh expansion of the Mallows' C_p criterion (using its prime spectrum) and through SA.[3]Similar solutions were drawn by both methods except that a prime spectrum energy analysis (0.9-quantile was taken) selected variable ♯4 (body mass index) as relevant instead of ♯7 (diabetes pedigree function), as the SA suggested.

TABLE III
SUCCESS RATE (SR[%]) AND SPECIFICITY (SP[%]) ON THE CLASSIFICATION OF THE DIABETES DATASET WITH AND WITHOUT IRRELEVANT FEATURES SELECTED THROUGH SENSITIVITY ANALYSIS (AAG MEASURE) AND PRIME SPECTRUM $\mathbf{B}'$ FOR TRAINING (T) AND VALIDATION (V). RESULTS OF THE ADAP ORIGINAL ALGORITHM ARE ALSO SHOWN FOR COMPARISON PURPOSES

Measures	**ADAP Algor.**	**MLP (8×2×1)**	**MLP(AAG) (5×3×1)**	**MLP(B′) (5×3×1)**
SR[%] (T)	-	74.95	78.28	79.55
SP[%] (T)	-	83.48	87.69	88.31
SR[%] (V)	76	77.04	74.32	75.40
SP[%] (V)	65	80.84	79.64	80.40

The MLP model was built, varying the number of hidden neurons (<15 to avoid overfitting), the weight initialization range, and the learning rate (between 0.001 and 0.3) to determine the best topology through the cross-validation method. We used 576 training instances and tested performance on the remaining 192 instances, for proper comparison with previous work [38]. Training was accomplished using the familiar *back-propagation* (BP) algorithm and models were selected through cross-validation.

Table III shows results obtained with a dedicated MLP for each of the three subsets; taking all of the available features and with the ones selected through SA and Walsh expansion (prime spectrum $\mathbf{B}'$). The optimal solution is achieved with an MLP when using all of the available features. As the input space is reduced, specificity (SP[%]) and success rates (SR[%]) increase in the training set but decrease in the validation set. However, the ranking of features provided by the prime spectrum of the Mallow's C_p criterion produces a good compromise between complexity and outcomes in the validation set, which suggests that a faithful feature selection has been accomplished.

VI. CONCLUSION

In this paper, we have presented two improvements for the genetic search of features in supervised classification problems. First, we have presented a novel genetic operator (*m-features* operator), which fixes the number of features selected by the GA to m features. The *m-features* operator reduces the size of the features search space, which improves the GA performance in terms convergence time, and may lead to find very high quality solutions. Second, we have proposed a modification of the spectrum of the GA fitness function in order to perform ranking of the problem's features.

We have applied the GA with the modifications to three real problems–the *Thrombin binding problem*, the problem of *risk assessment of unstable angina*, and the *diagnosis of diabetes mellitus*. In the first problem, we have used the GA as a wrapper approach, achieving very good performance of the GA with the

[3] Networks were retrained after each feature selection run, as proposed in [36] and [37]. This methodology ensures effectiveness in the feature selection process.

m-features operator. In the other problems, we have shown the application of our improvements to the GA used as filter method for the FSP. In this application, we have also shown that the modified spectrum is a good parameter for performing ranking of the problem's features.

ACKNOWLEDGMENT

The authors would like to thank Dr. G. Sanz and Dr. M. Cardona from the Hospital Clínic i Provincial de Barcelona and to J. Marrugat from the Institut Municipal de Investigació Mèdica, Barcelona. The authors are indebted to E. Soria-Olivas for his support on this work and his helpful comments and discussions on the angina problem.

REFERENCES

[1] H. Liu and H. Motoda, *Feature Extraction Construction and Selection: A Data Mining Perspective*. Norwell, MA: Kluwer, 1998.

[2] G. F. Hughes, "On the mean accuracy of statistical pattern recognizers," *IEEE Trans. Inform. Theory*, vol. IT-14, pp. 55–63, Jan. 1968.

[3] K. Fukunaga and R. R. Hayes, "Effects of sample size in classifier design," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 11, pp. 873–885, Aug. 1989.

[4] R. Kohavi and G. H. John, "Wrappers for features subset selection," *Int. J. Digit. Libr.*, vol. 1, pp. 108–121, 1997.

[5] J. Yang and V. Honavar, "Feature subset selection using a genetic algorithm," *IEEE Intell. Syst.*, vol. 13, pp. 44–49, Mar./Apr. 1998.

[6] J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio, and V. Vapnik, "Feature selection for SVMs," in *NIPS*, 2000, pp. 668–674.

[7] P. Leray and P. Gallinari, "Feature selection with neural networks," *Behaviormetrika*, vol. 26, Jan. 1999.

[8] B. Hassibi and D. G. Stork, "Second order derivatives for network pruning: optimal brain surgeon," in *Advances in Neural information Processing Systems*, S. J. Hanson, J. D. Cowan, and C. L. Giles, Eds. San Mateo, CA: Morgan Kaufmann, 1993, vol. 5, pp. 164–171.

[9] Y. LeCun, J. Denker, S. Solla, R. E. Howard, and L. D. Jackel, "Optimal brain damage," in *Advances in Neural Information Processing Systems II*, D. S. Touretzky, Ed. San Mateo, CA: Morgan Kauffman, 1990.

[10] V. Tresp, R. Neuneier, and H. G. Zimmermann, "Early brain damage," in *Advances in Neural Information Processing Systems*, M. C. Mozer, M. I. Jordan, and T. Petsche, Eds. Cambridge, MA: The MIT Press, 1997, vol. 9, p. 669.

[11] K. Torkkola and W. M. Campbell, "Mutual information in learning feature transformations," in *Proc. 17th Int. Conf. Mach. Learning*, 2000, pp. 1015–1022.

[12] K. Torkkola, "On feature extraction by mutual information maximization," in *Proc. IEEE Int. Conf. Acoustics, Speech Signal Processing*, vol. 1, 2002, pp. 821–824.

[13] L. Talavera, "Feature selection and incremental learning of probabilistic concept hierarchies," in *Proc. 17th Int. Conf. Mach. Learning*, 2000, pp. 951–958.

[14] I. T. Jolliffe, *Principal Component Analysis*. New York: Wiley, 1986.

[15] A. Hyvarinen, J. Karhunen, and A. Oja, *Independent Component Analysis*. New York: Wiley, 2001.

[16] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees*, 3rd ed. London, U.K.: Chapman & Hall, 1984.

[17] T. Kohonen, *Self-Organizing Maps*, 3rd extended ed. New York: Springer-Verlag, 2001, vol. 30, Springer Series in Information Sciences.

[18] T. E. Campos, I. Bloch, and R. M. Cesar Jr., "Feature selection based on fuzzy distances between clusters: first results on simulated data," *Lecture Notes in Computer Science*, vol. 2013, p. 186, 2001.

[19] A. Blum and P. Langley, "Selection of relevant features and examples in machine learning," *Artif. Intell.*, vol. 97, pp. 245–271, 1998.

[20] D. E. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning*. Reading, MA: Addison-Wesley, 1989.

[21] H. Y. Chen, T. C. Chen, D. Min, G. Fischer, and Y. M. Wu, "Prediction of tacrolimus blood levels by using the neural network with genetic algorithm in liver transplantation patients," *Therapeutic Drug Monitoring*, vol. 21, no. 1, pp. 50–56, Feb. 1999.

[22] J. Sepúlveda-Sanchis, G. Camps-Valls, E. Soria-Olivas, S. Salcedo-Sanz, C. Bousoño-Calzón, G. Sanz-Romero, and J. Marrugat de la Iglesia, "Support vector machines and genetic algorithms for detecting unstable angina," in *Proc. Int. Conf. Computers in Cardiology*, Memphis, TN, Sept. 2002, pp. 261–265.

[23] W. Hordijk and P. F. Stadler, "Amplitude spectra of fitness landscapes," *J. Complex Syst.*, vol. 1, no. 1, pp. 39–66, Nov. 1998.

[24] M. D. Vose and A. H. Wright, "The simple genetic algorithm and the Walsh transform: part I, theory," *Evol. Comput.*, vol. 6, no. 3, pp. 253–273, Mar. 1998.

[25] P. S. Bradley, U. M. Fayyad, and O. L. Mangasarian, "Mathematical programming for data mining: formulations and challenges," *INFORMS J. Comput.*, vol. 11, no. 3, pp. 217–238, 1999.

[26] G. John, R. Kohavi, and K. Pfleger, "Irrelevant features and the subset selection problem," in *Proc. 11th Int. Conf. Machine Learning*, Aug. 1994, pp. 121–129.

[27] S. Salcedo-Sanz, M. Prado-Cumplido, F. Pérez-Cruz, and C. Bousoño-Calzón, "Feature selection via genetic optimization," in *Proc. Int. Conf. Artificial Neural Networks*, Lecture Notes in Computer Science. Madrid, Spain, Aug. 2002, pp. 547–552.

[28] H. Kargupta, "A striking property of genetic code-like transformations," *J. Complex Syst.*, vol. 13, no. 1, pp. 1–32, 2000.

[29] KDDNuggets Inc.. (2001) Annual KDD Cup. [Online] Available: http://www.cs.wisc.edu/~-dpage/kddcup200l/

[30] J. Cheng, D. Bell, and W. Liu, "An algorithm for beyesian belief network construction from data," in *Proc. Nat. Acad. Sci.*, vol. 97, 1997, pp. 262–267.

[31] ——, "Learning belief networks from data: an information theory based approach," in *Proc. ACM Conf. Information Knowledge Management*, 1997.

[32] L. Serés, V. Valle, J. Marrugat, G. Sanz, R. Masiá, J. Lupón, A. Curós, J. Sala, L. Molina, and M. Pavesi, "Usefulness of hospital admission risk stratification for predicting nonfatal acute myocardial infarction or death six months later in unstable angina pectoris," *Am. J. Cordial.*, vol. 84, pp. 963–969, 1999.

[33] V. Vapnik, *Statistical Learning Theory*. New York: Wiley, 1998.

[34] Y. Lin, Y. Lee, and G. Wahba, "Support Vector Machines for Classification in Nonstandard Situations," Dept. Stat. TR 1016, Univ. Wisconsin-Madison, 2000.

[35] G. B. Orr and K. R. Müller, *Neural Networks: Tricks of the Trade*. Berlin, Heidenberg, Germany: Springer-Verlag, 1998.

[36] A. N. Refenes, A. Zapranis, and G. Francis, "Stock performance modeling using neural networks: a comparative study with regression models," *Neural Networks*, vol. 7, no. 2, pp. 375–388, 1994.

[37] W. S. Sarle. (2000) How to Measure Importance of Inputs?. [Online] Available: ftp://ftp.sas.com/pub/neural/importance.html

[38] J. W. Smith, J. E. Everhart, W. C. Dickson, W. C. Knowler, and R. S. Johannes, "Using the ADAP learning algorithm to forecast the onset of diabetes mellitus," in *Proc. IEEE Comput. Soc. Press. Symp. Computer Applications and Medical Care*, Apr. 21–24, 1988, pp. 261–265.

Sancho Salcedo-Sanz (S'00–M'03) was born in Madrid, Spain, in 1974. He received the B.S. degree in physics from Universidad Complutense de Madrid, Madrid, Spain, in 1998, and the Ph.D. degree in telecommunications engineering from Universidad Carlos III de Madrid in 2002.

Currently, he is a Research Fellow in the School of Computer Science of the University of Birmingham, Birmingham, U.K. His research interests include optimization in communications, hybrid algorithms, and neural networks. He has coauthored more than 15 international journals and conference papers in the field of GAs and hybrid algorithms.

Gustavo Camps-Valls received the Ph.D. degree in physics from the Universitat de València, València, Spain, in 2002.

Currently, he is an Associate Professor in the Department of Electronics Engineering at the Universitat de Valencia. His research interests include neural networks and kernel methods for hyperspectral data classification, health sciences, and safety-related areas.

Fernando Pérez-Cruz (S'97–M'00) was born in Sevilla, Spain, in 1973. He received the Bachelor's degree in telecommunications engineering from the Universidad de Sevilla, Seville, Spain, in 1996, and the Ph.D. degree in telecommunications engineering from the Universidad Politécnica de Madrid, Madrid, Spain, in 2000.

Currently, he is an Associate Professor in the Department of Signal Theory and Communications at the Universidad Carlos III de Madrid. His research interests include adaptive learning for nonlinear signal processing and communications and infrared detection. He has co-authored many contributions in international journals and conferences.

José Sepúlveda-Sanchis received the B.Sc., M.Sc., and Ph.D. degrees in physics from the University of Valencia, Valencia, Spain.

Currently, he is a Postdoctoral Researcher in the biochemistry department of Baylor College of Medicine, Houston, TX. He completed a postdoctoral year in Karolinska Institute, Stockholm, Sweden, conducting research in bioinformatics.

Carlos Bousoño-Calzón (M'95) received the B.S. and Ph.D. degrees in telecommunications engineering from the Universidad Politécnica de Madrid, Madrid, Spain, in 1992 and 1996, respectively.

Currently, he is an Associate Professor in the Department of Signal Theory and Communications at the Universidad Carlos III de Madrid, Madrid, Spain. His research interests include optimization in communications, GAs, and neural networks. He has co-authored many contributions in international journals and conference papers in these areas.