World Academy of Science, Engineering and Technology
International Journal of Educational and Pedagogical Sciences
Vol:10, No:6, 2016

Multi-Objective Evolutionary Computation Based Feature Selection Applied to Behaviour Assessment of Children

F. Jiménez, R. Jódar, M. Martín, G. Sánchez, G. Sciavicco

Open Science Index, Educational and Pedagogical Sciences Vol:10, No:6, 2016 waset.org/Publication/10004843

Abstract—Attribute or feature selection is one of the basic strategies to improve the performances of data classification tasks, and, at the same time, to reduce the complexity of classifiers, and it is a particularly fundamental one when the number of attributes is relatively high. Its application to unsupervised classification is restricted to a limited number of experiments in the literature. Evolutionary computation has already proven itself to be a very effective choice to consistently reduce the number of attributes towards a better classification rate and a simpler semantic interpretation of the inferred classifiers. We present a feature selection wrapper model composed by a multi-objective evolutionary algorithm, the clustering method Expectation-Maximization (EM), and the classifier C4.5 for the unsupervised classification of data extracted from a psychological test named BASC-II (Behavior Assessment System for Children - II ed.) with two objectives: Maximizing the likelihood of the clustering model and maximizing the accuracy of the obtained classifier. We present a methodology to integrate feature selection for unsupervised classification, model evaluation, decision making (to choose the most satisfactory model according to a *a posteriori* process in a multi-objective context), and testing. We compare the performance of the classifier obtained by the multi-objective evolutionary algorithms ENORA and NSGA-II, and the best solution is then validated by the psychologists that collected the data.

Keywords—Feature selection, multi-objective evolutionary computation, unsupervised classification, behavior assessment system for children.

I. Introduction

The Behavior Assessment System for Children - II ed. (BASC-II) is a norm referenced diagnostic tool designed to assess the behavior and self-perceptions of children and young adults with ages from 3 to 18. The BASC-II is a multi-dimensional and multi-method tool, since it measures numerous behavioral and personality characteristics through several report-based measures. It can be used, among other objectives, for program planning, evaluation, and intervention, to determine educational classification and programming assistance eligibility, and to assist in determining the causes of behavioral problems for children with disabilities [1], [2]. The test is built over 149 questions (each referred to as Item), and it was designed to give professionals a single test that provides a global view of both the observable conduct as well as the self-perceived emotions of the subjects. The components of the test are focused on both clinical and adaptive aspects of the behaviour and the personality of the subjects, which allows to diagnose potential problems in these areas. For this experiments, data were collected from the administration of the BASC-II in the Spanish version [2] to 157 subjects, all scholars from the local elementary school *Colegio San Buenaventura*, located in Murcia (in south-eastern Spain). In this particular case, the BASC-II test was enriched with three more questions for statistical purposes: age (6, 7 or 8 years old), sex, and class (1 or 2).

The structure of the collected data from applying BASC-II, that is, a relatively high number of categorical features for a relatively limited number of subjects, presents an ideal environment for *data mining* processes (DM) that include a feature selection phase. DM [3] consists of applying algorithms to collected data for distinct purposes such as finding patterns, creating prediction models or obtaining statistical data; besides, as the amount of collected features grows, the performance of such algorithms becomes more and more critical. We are interested here in *unsupervised data classification* [3], that is, classification from a data set in which instances are not labeled with any class. In particular, we want to apply a *feature selection* mechanism with the goal of answering the following questions:

(i) Given the high number of features that are collected for each subject, is there a hidden knowledge that we can extract by selecting a subset of the features and transcending the specific dimensions of BASC-II to which they belong?

(ii) Can we apply to the selected features a clustering algorithm and obtain clusters that can be semantically interpreted?

(iii) Can we then classify the subjects with respect to the selected features and the obtained labels?

Feature selection is an independent process whose main objective is to reduce the dimension of the data set (i.e., the number of input variables) in order to perform a more efficient and more easily interpretable classification, which is also more accurate because the noise introduced by irrelevant features has been removed. Feature selection methods (see [4] for a survey) roughly range from *filter models* to *wrapper models*, to *embedded models*, according to their level of interaction with the learning algorithm. Feature selection has become increasingly frequent in classification or regression applications in genomics, health sciences, economics, finance,

Maria del Pilar Martín and Rosalía Jódar are with the Faculty of Psychology, University of Murcia, 30100 Espinardo, Murcia, Spain (e-mail: mpmartin@um.es, rosalia.jodar@um.es).

Fernando Jiménez, Gracia Sánchez, and Guido Sciavicco are with the Faculty of Computer Science, University of Murcia, 30100 Espinardo, Murcia, Spain (e-mail: fernan@um.es, gracia@um.es, guido@um.es).

Open Science Index, Educational and Pedagogical Sciences Vol:10, No:6, 2016 waset.org/Publication/10004843

among others (see, e.g., [5], [6]), as well as in psychology and social sciences (see, e.g., [7], [8]).

In this paper, we apply a wrapper feature selection mechanism via the adaptation of the multi-objective evolutionary algorithms known as ENORA [9], [10] and NSGA-II [11] with two objectives:

(i) Maximizing the likelihood of the cluster (via the Expectation-Maximization algorithm);

(ii) Maximizing the accuracy of classifier (via the C4.5 algorithm).

We compare the performances of ENORA and NSGA-II for this task by measuring the quality of the classifiers that have been built over the selected features. The most relevant characteristic of this particular experiment is the interaction among the evolutionary algorithm(s), the clustering method, and the classification algorithm, which naturally leads to a Pareto-optimal final population that requires a *a posteriori* decision method, which we have devised and which returned excellent results.

II. Background

(Un)supervised learning in medicine and psychology: *Cluster analysis* and *unsupervised classification* were initially used within the disciplines of biology and ecology [12]. Although these techniques have been employed in the social sciences, they have not gained the same widespread popularity as in the natural sciences. A general interest in cluster analysis increased in the 1960s, resulting in the development of several new algorithms that expanded possibilities of analysis. It was during this period that researchers began utilizing various innovative tools in their statistical analyses to uncover underlying structures in data sets. Within a decade, the growth of cluster analysis and its algorithms reached a high point. By the 1970s, the focus shifted to integrating multiple algorithms to form a cohesive clustering protocol [13]. In recent decades, there has been a gradual incorporation of cluster analysis into other areas, such as the health and social sciences. However, the use of cluster analysis within the field of psychology continues to be infrequent [14]. A recent survey on the application of unsupervised classification to psychological data can be found in [15], and clustering and classification methods to psychology are also described in [16]. On the other hand, (supervised) *classification* has been extensively applied in the general field of medicine. One illustrative milestone is the MYCIN system, a diagnosis support system for infectious diseases in which the medical knowledge is provided from the physician's team in the form of rules [17]. Classification based on fuzzy rules has been also applied in the classification of medical images [18], interpretation of mammograms [19], and survival prediction in burn patients [10].

Feature selection: *Feature selection* [20] is the process of eliminating features from the data set that are irrelevant with respect to the task to be performed. Its main aim is to determine a minimal subset of features from a problem domain while retaining a suitably high accuracy in representing the original features. Feature selection finds useful features to represent the data and remove non-relevant ones, and simplifies the implementation of the classifier itself by determining what features should be made available to it. Furthermore, feature selection tends to speed up the processing rate of the classifier; at the same time, it improves response times by reducing the dimensionality of the input space. Additionally, feature selection can improve the quality of the classification in terms of accuracy and interpretability of the outcome. According to whether the training set is labelled (classified) or not, feature selection algorithms can be categorized into *supervised* and *unsupervised*, respectively. In between it lies a third category, that is, *semi-supervised* feature selection, which includes algorithms that make use of both labeled and unlabelled data to estimate the relevance of a feature. Feature selection methods can be further categorized into *filter models*, *wrapper models* and *embedded models*. The filter model separates feature selection from classifier learning, so that the learning algorithm does not interact with the selection algorithm; features may be ranked independently of the feature space (*univariate scheme*), or they may be evaluated in batch (*multivariate scheme*), being in this way naturally capable of handling redundancy. Filter techniques easily scale to very high-dimensional data sets, they are computationally simple and fast, and they are independent of the classification algorithm. Filter methods, however, ignore the interaction with the classifier. On the other hand, the wrapper model is based on the predictive accuracy of a predetermined learning algorithm to determine the quality of selected features, so that for each selection the learning algorithm is asked to check the accuracy of the classifier built over it. Finally, the embedded model selects the best features according to accuracy while building the model, and the interaction between selection algorithm and learning algorithm is integrated step-by-step. According to the type of the output, feature selection algorithms can be classified into *feature weighting* algorithms and *subset selection* algorithms: feature selection algorithms in filter and embedded models may return either a subset of selected features or weights that measure the relevance of each feature. On the other hand, feature selection algorithms with wrapper models usually return feature subsets, and are therefore classified as subset selection algorithms. Feature selection methods typically consist of four basic steps, namely, *generation*, *evaluation*, *stopping criterion*, and *validation*. The generation phase heuristically searches in the entire space whose dimension is 2^N, where N is the number of features; a candidate feature subset is chosen based on a given search strategy, and sent, in the second step, to be evaluated according to a certain evaluation criterion. The subset that best fits the evaluation criterion is chosen among all the candidates that have been evaluated after the stopping criterion are met. In the final step, the chosen subset is validated using domain knowledge or a validation set. Examples of subset generation schemata include, among others, *greedy hill-climbing approach* [5], *sequential forward selection* [21], *sequential backward elimination* [22], and *bi-directional selection* [20] (see [4] for a survey). In this paper, we present the use of a *random* search method, as in [23], [24], [25], and, in particular, of a multi-objective evolutionary algorithm.

Multi-objective evolutionary feature selection: In the evolutionary computational model, a problem plays the role of an environment populated by a set of individuals, each one representing a possible solution to the problem. The degree of adaptation of each individual to its environment is expressed by an adequacy measure known as a *fitness function*. Starting with an initial population of random solutions, in each iteration the best individuals are selected and combined using variation operators such as *crossing* and *mutation* to build the next generation. This process is repeated until some stop criterion is met, typically based on the number of iterations. The use of evolutionary strategies for the selection of features has been initially proposed in [26]. Since then, it has been regarded as a powerful tool for feature selection in machine learning [24] and proposed by numerous authors as a search strategy (see, e.g., [27], [28]). *Multi-objective* evolutionary algorithms are designed to solve a set of minimization/maximization problems for a tuple of n functions $f_1(\overrightarrow{x}), \ldots, f_n(\overrightarrow{x})$, where $\overrightarrow{x}$ is a vector of parameters belonging to a given domain. Let $\mathcal{F}$ be the search space for a multi-objective optimization problem. A solution $\overrightarrow{x} \in \mathcal{F}$ is said to be a *non dominated* (or *Pareto optimal*) if and only if there exists no $\overrightarrow{y} \in \mathcal{F}$ for which: *(i)* there exists $1 \leq i \leq n$ such that $f_i(\overrightarrow{y})$ improves $f_i(\overrightarrow{x})$, and *(ii)* for each $j \neq i$, $f_j(\overrightarrow{x})$ does not improve $f_i(\overrightarrow{y})$.The set of non dominated solutions from $\mathcal{F}$ is called *Pareto front*.

Multi-objective approaches are particularly suitable for multi-objective optimization, as they search for multiple optimal solutions in parallel; such algorithms are capable of finding a set of optimal solutions in its final population in a single run, and once the set of optimal solutions is available, the most satisfactory one can be chosen by applying a preference criterion. In subset feature selection, each solution in the Pareto front represents a subset of features with an associated trade-off between, for example, accuracy and data set dimension.

In the first evolutionary approach involving multi-objective feature selection [29], three criteria (accuracy, number of features and number of instances) are aggregated and then a single-objective optimization algorithm is applied. A formulation of feature selection as a multi-objective optimisation problem using a neuro-fuzzy based wrapper has been proposed in [30]. Other approaches, such as [31], [32], [33], [34], [35], propose the use of NSGA-II [11] in combination with wrapper methods that use decision tree (such as C4.5 [34]), support vector machines [32], [33], maximal entropy based models [31], or a filter method [35] that include measures of consistency, dependency and distance information.

The Expectation-Maximization algorithm: The ExpectationMaximization (EM) algorithm [36] is an iterative method for finding maximum likelihood estimates of parameters in statistical models, where the model depends on unobserved latent variables. An iteration alternates between performing an *expectation* (E) step, which creates a function for the expectation of the likelihood evaluated using the current estimate for the parameters, and a *maximization* (M) step, which computes parameters maximizing the expected likelihood found on the E step. These parameter-estimates are then used to determine the distribution of the latent variables in the next E step. When used for clustering (unsupervised learning) the E-step estimates the distribution of labels, and the M-step chooses new parameters to maximize the expected likelihood of the observed data. The EM algorithm is available in Weka [37].

The classifier J48: J48 is the Weka [37] implementation of the decision tree C4.5 introduced in [38] (as an improvement of algorithm ID3, by the same author). It is known to be computationally very efficient and to guarantee the interpretability of the results. Briefly, C4.5 builds decision trees from a set of training data by using the *information entropy gain* criterion. At each node of the tree, C4.5 chooses the attribute of the data that most effectively splits its set of samples into subsets, each one belonging to one of the predefined classes. The splitting criterion is the *normalized information gain*: the feature with the highest normalized information gain is chosen to make the decision.

Discussion: As we have seen, numerous approaches that use multi-objective evolutionary algorithms for feature selection have been proposed in recent years. Although both filters as wrapper methods have been proposed, most of the authors use subset selection wrapper methods. The optimization model most commonly used has been maximizing the accuracy of the classifier along with minimizing the number of features (see also [39]), although many other models have been proposed for specific contexts. We propose a wrapper-based subset evaluation strategy for feature selection in unsupervised learning based on a multi-objective algorithm (ENORA or NSGA-II), with two objectives: maximizing the likelihood of the clusterization (via the EM algorithm) and the accuracy of the classification (via the C4.5 algorithm).

III. ENORA: AN ELITIST PARETO-BASED MULTI-OBJECTIVE EVOLUTIONARY ALGORITHM

While NSGA-II is a standard multi-objective evolutionary algorithm [40] widely used in wrapper-based feature selection, applying ENORA [9], [10], [39] to this task is relatively new. Its main components, that is, representation, fitness functions, initial population, selection and sampling mechanisms, generational replacement schemata, and variation operators are briefly described in this section.

The ENORA algorithm: ENORA is an elitist Pareto-based multi-objective evolutionary algorithm that uses a $(\mu + \lambda)$ survival, where μ corresponds to the population size and λ refers to the number of generated children. The $(\mu + \lambda)$ strategy was originally developed in [41] as an *evolution strategy*, using selection, adapting mutation and a population of size one, called $(1 + 1)$-ES. Recombination and populations with more than one individual were later introduced in [42]. The $(\mu + \lambda)$ technique allows the μ best children and parents to survive and is, therefore, an elitist method. ENORA uses a $(\mu + \lambda)$ survival with $\mu = \lambda$, where μ and λ are equal to the population size, binary tournament selection, and self-adaptive crossover and mutation for multi-objective evolutionary optimization (Algorithm 1).

After the initialization and evaluation of a population P of N individuals, and for each of the T generations, a pair

Algorithm 1 $(\mu + \lambda)$ strategy for multi-objective optimization

```
Require: $T > 1$ {Number of iterations}
Require: $N > 1$ {Number of individuals in population}
Initialize $P$ with $N$ individuals
Evaluate all individuals of $P$
$t \leftarrow 0$
while $t < T$ do
    $Q \leftarrow \emptyset$
    $i \leftarrow 0$
    while $i < N$ do
        $Parent1 \leftarrow$ Binary tournament selection from $P$
        $Parent2 \leftarrow$ Binary tournament selection from $P$
        $Child1, Child2 \leftarrow$ Self-adaptative variation $Parent1, Parent2$
        Evaluate $Child1$
        Evaluate $Child2$
        $Q \leftarrow Q \bigcup \{Child1, Child2\}$
        $i \leftarrow i + 2$
    end while
    $R \leftarrow P \bigcup Q$
    $P \leftarrow N$ Best individuals from $R$ according to the Rank-crowding better
    function in population $R$
    $t \leftarrow t + 1$
end while
return Non dominated individuals from $P$
```

Algorithm 2 Binary tournament selection

```
Require: $P$ {Population}
$I \leftarrow$ Random selection from $P$
$J \leftarrow$ Random selection from $P$
if $I$ is better than $J$ according to the rank crowding better function in population $P$ then
    return $I$
else
    return $J$
end if
```

Algorithm 3 Rank-Crowding-Better function

```
Require: $P$ {Population}
Require: $I, J$ {Individuals to compare}
if $rank\,(P, I) < rank\,(P, J)$ then
    return $True$
end if
if $rank\,(P, J) < rank\,(P, I)$ then
    return $False$
end if
return $crowding\_distance\left(P^I, I\right) > crowding\_distance\left(P^J, J\right)$
```

of parents are selected by a *binary tournament selection* from the population P (Algorithm 2). This algorithm returns the best between two random individuals according to the *rank crowding better function* (Algorithm 3). An individual I is considered better than an individual J if the rank of individual I is better (lower) than the rank of individual J in the population P. The *rank* of an individual I in a population P, $rank\,(P, I)$, is the *non-domination level* of the individual I among the individuals J of the population P so that $slot\,(I)$=$slot\,(J)$, where the radial *slot* ($slot\,(I)$) is the portion of the search space to which I belongs, and it is defined as:

$$slot\,(I) = \sum_{j=1}^{n-1} d^{j-1} \lfloor d \frac{\alpha_j^I}{\pi/2} \rfloor \tag{1}$$

$$\alpha_j^I = \begin{cases} \frac{\pi}{2} & \text{if } h_j^I = 0 \\ \arctan(\frac{h_{j+1}^I}{h_j^I}) & \text{if } h_j^I \neq 0 \end{cases} \tag{2}$$

where $d = \left\lfloor \sqrt[n-1]{N} \right\rfloor$ and h_j^I is the objective function f_j^I normalized in $[0, 1]$. If two individuals I and J have the same rank, the best individual is the individual with the greater crowding distance in its front (the front of I and J are denoted by P^I and P^J, respectively, Algorithm 3). The selected pair of parents is crossed, mutated, evaluated and added to an initially empty auxiliary population Q. This process is repeated until Q contains a number N of individuals. An auxiliary population R is obtained with the union of the populations P and Q. Then, the rank of all individuals in the population R is calculated (Algorithm 3). Finally, the N best individuals of R according to the rank crowding better function (Algorithm 3) survive to the next generation.

The crowding distance of an individual I in a population P is a measure of the search space around I which is not occupied by any other individual in the population P. This quantity serves as an estimate of the perimeter of the cuboid formed by using the nearest neighbours as the vertices. If we define $f_j^{max} = \max_{I \in P} \left\{ f_j^I \right\}$, $f_j^{min} = \min_{I \in P} \left\{ f_j^I \right\}$, and $f_j^{sup_j^I}$ (resp., $f_j^{inf_j^I}$) is the value of the jth objective function for the individual higher adjacent (resp., lower adjacent) in the jth objective function to the individual I, then the crowding distance $crowding\_distance(P, I)$ is ∞ if for each j it is the case that $f_j^I = f_j^{max}$ or $f_j^I = f_j^{min}$, and it is

$$crowding\_distance\,(P, I) = \sum_{j=1}^{n} \frac{f_j^{sup_j^I} - f_j^{inf_j^I}}{f_j^{max} - f_j^{min}} \tag{3}$$

otherwise.

Representation, evaluation and variation: We use a fixed-length representation where each individual consists of a bit set, and each bit represents a selected (1) or non selected (0) feature. Additionally, to carry out self-adaptive crossing and mutation, each individual has two discrete parameters $d_I \in \{0, \ldots, \delta\}$ and $e_I \in \{0, \ldots, \epsilon\}$ associated to, respectively, crossing and mutation, where $\delta \geq 0$ is the number of crossing operators and $\varepsilon \geq 0$ is the number of mutation operators. Therefore, an individual I in the feature selection problem with M features is represented as:

$$I = \left\{ b_1^I, \ldots, b_M^I, d_I, e_I \right\}$$

where for each i $b_i^i \in \{0, 1\}$, and where $d_I \in \{0, \ldots, \delta\}$, $e_I \in \{0, \ldots, \epsilon\}$. An individual I is evaluated with two fitness functions, $f_1\,(I)$ and $f_2\,(I)$, corresponding to the two objectives of the multi-objective optimization model:

$$\begin{cases} f_1\,(I) = \mathcal{ACC}\,(I) \\ f_2\,(I) = \mathcal{LIKE}\,(I) \end{cases} \tag{4}$$

The quantity $\mathcal{ACC}\,(I)$ is defined as

$$\mathcal{ACC}\,(I) = \frac{N_c}{N_t}, \tag{5}$$

where N_c and N_t are the number of correctly classified instances and the number of total instances, respectively, and it is the *accuracy* of the classifier, when classification is performed using only the attributes in individual I: $f_1\,(I)$ must be maximized. The quantity $\mathcal{LIKE}\,(I)$ is the *log likelihood* of the clustering model obtained by the EM algorithm. This is

Algorithm 4 Variation

```
Require: $Parent1$, $Parent2$ {Individuals to vary}
$Child1 \leftarrow Parent1$
$Child2 \leftarrow Parent2$
Self-adaptive crossover $Child1$, $Child2$
Self-adaptive mutation $Child1$
Self-adaptive mutation $Child2$
return $Child1, Child2$
```

Algorithm 5 Adaptive crossover

```
Require: $I$, $J$ {Individuals to cross}
Require: $p_v$ $(0 < p_v < 1)$ {Probability of operator change}
Require: $\delta > 0$ {Number of different crossover operators ($\delta = 1$ in our case)}
if A random Bernoulli variable of probability $p_v$ takes the value 1 then
    $d_I \leftarrow$ Int Random from $\{0, \delta\}$
end if
$d_J \leftarrow d_I$
Carry out the type of crossover specified by $d_I$:
    {0: No cross}
    {1: Uniform crossover}
```

obtained via an initial estimation of the unknown parameter Z (the class) and then iteratively maximizing the value of $log\ p(X, Z|\Theta)$ by computing the parameter Θ; intuitively, X is the set of subjects restricted to the selected features, and interpreted as statistical variables, Z is the class variable, and Θ is set of Gaussian parameters that must be found. Notice that the EM algorithm works by estimating both the number of classes and their centroids, and that the logarithm function is used to simplify the calculation [43], [36]. The function $f_2(I)$ must be maximized. Finally, unlike similar wrapper-based approaches, the cardinality of the subset of the selected features is not constrained as an objective, as there is no direct relation between the number of features and the likelihood of the clusters.

The initial population is generated randomly. For each individual I in the population, q randomly chosen bits are set to 1, and the remaining $M - q$ to 0 (in this way, we ensure the diversity of the initial population); moreover, the parameters d_I and e_I are also randomly generated in their respective domains $\{0, \delta\}$ and $\{0, \epsilon\}$. Self-adaptive crossover and mutation are used to maintain diversity in the population and to sustain the convergence capacity of the evolutionary algorithm. In a self-adaptive evolutionary algorithm [44], the probabilities of crossover and mutation vary according to the fitness value of the solutions. By using self-adaptive variation operators, it is not necessary to set the probabilities of the application of the different operators *a priori*. We use *uniform crossover* and *one flip mutation*, although other variation operators may be considered. The selection of the operators is made by means of the adaptive technique according to the parameters d_I and e_I that indicate which crossover and mutation is carried out for individual I.

IV. Experiment Design and Results

In this section, we present the results of the experiment over our data set, obtained by a methodology which includes pre-processing of the data, feature selection, optimizers' performances comparison (based on hypervolume metrics), classifier learning construction, and test, as shown in Fig. 1.

Algorithm 6 Adaptive mutation

```
Require: $I$ {Individual to mutate}
Require: $p_v$ $(0 < p_v < 1)$ {Probability of operator change}
Require: $\epsilon > 0$ {Number of different mutation operators ($\epsilon = 1$ in our case)}
if A random Bernoulli variable of probability $p_v$ takes the value 1 then
    $e_I \leftarrow$ Int Random from $\{0, \epsilon\}$
end if
Carry out the type of mutation specified by $e_I$:
    {0: No mutation}
    {1: One flip mutation}
```

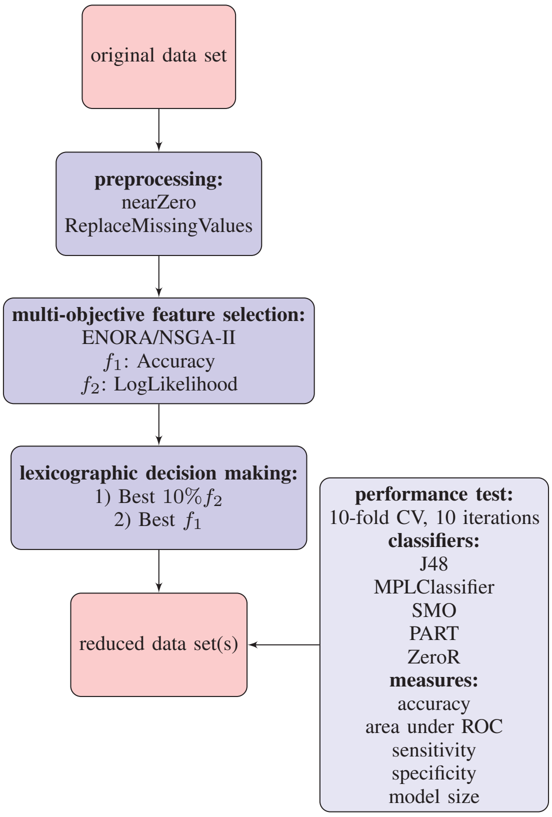

Fig. 1 Proposed methodology

The BASC-II data set: The BASC-II [1], [2] test is composed by 149 questions (each referred to as Item), to which three more questions have been added in this specific case for statistical purposes (namely, age, sex, and class); the teachers' version T2 [2] was selected, so that subjects were tested by means of a questionnaire filled in by their teachers. Each question is categorical, with possible answers from 1 to 4 (from *fully disagree* to *fully agree*). The 149 original questions can be categorized into two *dimensions*, namely the *clinical dimension* and the *adaptive dimension*; the clinical dimension focuses on *problems' exteriorization*, *problems' internalization*, and *school-related problems*, so that the whole set of questions contains in fact questions of four (sub)dimensions. A total of 157 subject, all children between 6 and 8 years old have been considered for this test, whose original purpose was to obtain a detailed history of their social, psychological, and educational development, in order to classify their observed behaviour, and, possibly, to design a specific intervention plan. BASC-II is well-known and accepted in the psychology community, and we want to apply a wrapper feature selection model in order to establish whether it is possible to classify a subject based on his/her answer to the BASC-II questions before its semantical interpretation.

Data pre-processing: The initial data set composed of 152 features has been pre-processed as follows. First, we have replaced all the missing values with the modes of the attributes; to this end, the procedure *ReplaceMissingValues* from the *weka.filters.unsupervised.attribute* package has been used. Second, we have eliminated the features with too small variation; we have used the procedure *NearZeroVar* from Caret R [45] for this task. There were 32 features that presented a zero or near-to-zero variance, and all of them referred to very extreme behaviours (clinical alterations), highly unlikely to be found in standard populations. As an example of eliminated features, Item 14 is *He/she has sphincters' control problems*: for children between 6 and 8 years old this is a problem that only presents itself in association to a physiological condition, or to a phycological condition connected with the improper internalisation of a negative situation. Similarly, Item 32, that is, *He/she coerces and intimidates others*, has been eliminated as well: as a matter of fact, such an extreme behaviour in minors is only found in association to some kind of pathology.

Feature selection: Both search strategies ENORA and NSGA-II have been integrated into a wrapper feature selection method based on C4.5 and EM, using the two objective functions described in the previous section: accuracy maximization and likelihood maximization. After 30 runs, to each non-dominated individual of the last population of each strategy, we performed a 10-folds cross-validation to each non-dominated solution of the last population under the accuracy and the (log)likelihood parameters. More in particular, over the set of non-dominate solutions we selected the best 10% w.r.t. the (log)likelihood; of these, we selected the one with best accuracy (one for ENORA, and one for NSGA-II). Finally, for each of the two solutions, we built a reduced data set based on the the selected features. The two search strategies have been implemented with dynamically adapted parameters [44], and written in Java by using the Weka [37] package. For each run, with respect to the classification, we used the following evaluator: *weka.attributeSelection.WrapperSubsetEval -B weka.classifiers.trees.J48 -F 5 -T 0.01 -R 1 -E acc – -C 0.25 -M 2*. Both search strategies ENORA and NSGA-II have been run with the number of evaluations set to 100 and the number of individuals in each population set to 100, for a total of 10000 evaluations. The search strategy ENORA has been officially incorporated by authors into Weka under the identifier *MultiObjectiveEvolutionarySearch*.

Performance test: Since the search for an optimal subset of features is performed by two different evolutionary algorithms, it makes sense to compare the hypervolume of the two executions, that is, the volume of the search space dominated by a population P [40]. While the purpose of this experiment is not to establish which evolutionary algorithm behaves better for this task, the statistical comparison of the results obtained by the two optimizers, a confidence interval of 90% has been used for the mean obtained with a pairwise t-test [46] is shown in Table I. From it, and from the respective box plots (Fig. 2), we deduce that NSGA-II shows a slightly better behaviour, although not statistically significant. Moreover, in terms of the performance of the obtained classifier, as we discuss below, the data set obtained by ENORA produces better results.

TABLE I
STATISTICS FOR THE HYPERVOLUME OBTAINED WITH 30 RUNS

	ENORA	NSGA-II
Minimum	0.2963	0.2297
Maximum	0.5322	0.5119
Mean	0.3850	0.3277
S.D.	0.0527	0.0571
C.I. Low	0.3653	0.3063
C.I. High	0.4047	0.3490

S.D = Standard Deviation of Mean
C.I. = Confidence Interval for the Mean (95%)

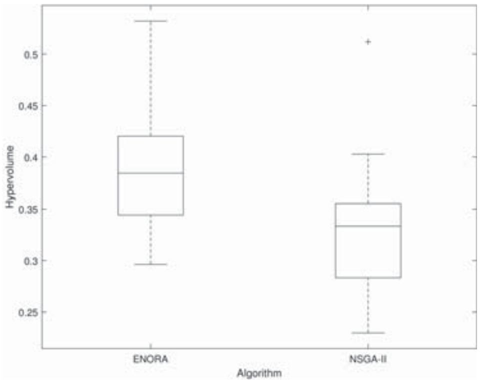

Fig. 2 Hypervolume boxplots: ENORA against NSGA-II

Once the two reduced data sets (ENORA-DS and NSGA-II-DS) have been obtained, we tested and compared them. First of all, notice that EM found three clusters when run both with ENORA and NSGA-II, but one of them presented always with no subjects in both cases. This means that the subject can be classified into subjects C1 and subjects C2 thanks to their performance on BASC-II. ENORA selected 10 features, and NSGA-II 9; 6 features were selected by both methods (see Table II). We configured the *Experimenter* tool available in Weka with the two data sets to perform a 10-fold cross-validation (10 iterations) with the following classifiers: J48, MLPClassifier (which trains a *multi-layer perceptron* with one hidden layer using Weka's Optimization class, by minimizing the squared error plus a quadratic penalty with the so-called *BFGS method*), SMO (which implements Platt's sequential minimal optimization algorithm for training a support vector classifier [47]), PART (which produces a decision list from partial C4.5 decision tree by making the best leaf into a rule [48]), and ZeroR (which is a simple classifier to predict the mode for a nominal class [48]), all run with default parameters set by Weka. To analyze the result of the experiment we performed a paired t-test *corrected*, with 0.05 significance (being ENORA-DS the test base) and the following measures have been compared:*(i)* the percent of correct classifications; *(ii)* the (weighted) area under the ROC curve; *(iii)* the sensitivity (the true positive rate); *(iv)* the specificity (the false positive rate); *(v)* the serialized model

TABLE II
SELECTED QUESTIONS

Item	Question	ENORA-DS	NSGA-II
22	He/She makes mistakes due to short attention spam	X	
35	He/She carefully analyzes a problem before solving it		X
41	He/She is often punished at school		X
49	He/She plays alone	X	
68	He/She criticizes others		X
72	He/She easily adapts him/her self to changes in everyday routine	X	
73	He/She name calls other children	X	X
83	He/She complaints of pains	X	X
99	He/She makes fun of others	X	X
110	He/She falls ill before important tests	X	X
119	He/She has headache	X	X
138	He/She has sight problems	X	X
143	He/She shows interests for others' ideas	X	

TABLE III
COMPARATIVE RESULTS OF THE TEST

	ENORA-DS	NSGA-II-DS	ORIGINAL-DS
		percent correct	
trees.J48	**96.04(4.91)**	95.33(5.28)	83.98(7.48)*
functions.MLPClassifier	99.91(1.05)	**99.94(0.82)**	93.39(5.80)*
functions.LibSVM	92.63(5.76)	94.16(5.09)	**95.72(5.04)**
rules.PART	96.54(4.58)	**97.02(5.13)**	87.60(7.66)*
rules.ZeroR	69.46(2.04)	69.46(2.04)^v	**65.63(2.68)***
		Weighted avg. area under ROC	
trees.J48	**0.96(0.06)**	**0.96(0.06)**	0.83(0.10)*
functions.MLPClassifier	**1.00(0.00)**	**1.00(0.00)**	0.99(0.02)
functions.LibSVM	0.88(0.09)	0.91(0.08)	**0.95(0.06)**
rules.PART	0.97(0.05)	**0.99(0.04)**	0.85(0.10)*
rules.ZeroR	**0.50(0.00)**	0.50(0.00)	**0.50(0.00)**
		True positive rate (sensitivity)	
trees.J48	**0.98(0.03)**	**0.98(0.03)**	0.84(0.07)*
functions.MLPClassifier	**1.00(0.01)**	**1.00(0.01)**	0.93(0.05)*
functions.LibSVM	0.93(0.06)	0.94(0.05)	**0.96(0.05)**
rules.PART	0.98(0.03)	**0.99(0.03)**	0.87(0.08)*
rules.ZeroR	**0.69(0.02)**	**0.69(0.02)**	0.66(0.03)*
		False positive rate (1-specificity)	
trees.J48	**0.05(0.08)**	**0.05(0.08)**	0.20(0.11)*
functions.MLPClassifier	**0.00(0.02)**	**0.00(0.02)**	0.09(0.08)*
functions.LibSVM	0.17(0.13)	0.13(0.11)	**0.06(0.08)**^{v}
rules.PART	**0.04(0.07)**	**0.01(0.06)**	0.17(0.12)*
rules.ZeroR	**0.69(0.02)**	**0.69(0.02)**	0.66(0.03)^v
		Serialized model size	
trees.J48	5209.39(384.05)	**5204.52(369.92)**	11430.69(464.73)^v
functions.MLPClassifier	**1400.00(0.00)**	11148.00(0.00)*	36734.00(0.00)^v
functions.LibSVM	**9995.12(494.18)**	16192.60(378.71)^v	156479.04(4883.08)*
rules.PART	**7419.37(1144.27)**	9480.65(810.80)*	13331.18(833.82)*
rules.ZeroR	**880.00(0.00)**	**880.00(0.00)**	846.00(0.00)^v

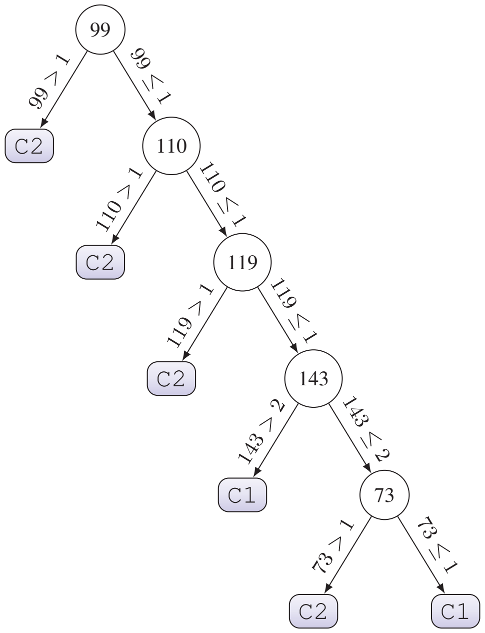

Fig. 3 Decision Tree from ENORA-DS

size.

Analysis of the solutions and discussion: Here we analyze the obtained solution based on the result of the above test, shown in Table III, where by ORIGINAL-DS we denote the original data set, after the pre-processing, and the execution of EM on all features. For each result, a mark * denotes that the result is statistically worse than the test base (ENORA-DS); similarly, a mark ^{v} denotes a statistically better result, and no mark denotes no statistically meaningful difference. The values between brackets are the standard deviations, and the boldfaced results are the best ones.

In terms of the performances of the classification model, we observe that: both feature selection methods were able to reduce the number of features in a very significative way, allowing an easier interpretation of the results, and the performances of all classifiers obtained by ENORA in terms of area under the ROC curve stands out; in particular, notice that the classifier obtained by MPLClassifier shows the perfect area (1.00). All classifiers obtained by ENORA and NSGA-II are significantly more sensitive than the ones built on the original data set, expect in one case.

In terms of result interpretation, consider, the decision tree extracted from ENORA-DS shown in Fig. 3. Only 5 out of 10 selected features were used in the tree, indicating that the remaining 5 are necessary for clusterization, but not for classification. The BASC-II test is designed to asses several different aspects of children's behaviour. Selecting a subset of such features allowed us to identify one particular aspect to which focus the attention in diagnosis, and, in particular, on the possession of certain *social-emotional* abilities. These are particularly interesting in the context of educational environments; social-emotional abilities [49], [50], [51] allow one to prevent risky conducts in children and provide children with tools for conflict and problem resolution, self-control, leadership, responsible decision-making, self-sufficiency, self-esteem, and self-awareness improvement, emotion managing and behaviour, relational abilities, among others. Social-emotional abilities and competencies play a role in several areas [52], such as sexual behaviour, social-moral cognition, problem solving, and academic performance. They improve the relationship with teachers, decrease the tendency to violent and aggressive behaviours, the probability of drug abuse in the adolescent age, and the improve risk management.

Subjects that belong to cluster C1 can be interpreted as those possessing certain social-emotional abilities, as *showing interest on others' ideas* (implying empathy and pro-social conducts), and not presenting a disruptive behaviour toward the peers (which shows self-control, the ability of managing the emotions, and propensity to social relationships). It can be concluded that C1 subjects posses a certain degree of ability to satisfactorily interact with peers. On the other hand, subjects belonging to cluster C2 show the absence of such social-emotional abilities. This indicates that such subjects may have troubles with emotional self-regulation, showing symptoms of somatization such as *headaches* or *falling ill before important tests*; they also present a certain degree of disruptive behaviour towards the peers, such as *name calling* or *making fun* of classmates (indicating absence of emotional competencies, troubles with inter-personal relationships, poor academic performance, violent or aggressive behaviour). A social-emotional reinforcement in these subjects is considered necessary.

V. Conclusions

In this paper, we considered a data set taken from the administration of the Behavior Assessment System for Children Test (BASC-II) to 157 subjects. Using a novel methodology based on feature selection via multi-objective evolutionary algorithms, the decision tree learning algorithm C4.5, and the Expectation-Maximization, we ran a wrapper method to maximize the accuracy of the classification as well as the likelihood of the clusterization. As a result, we were able to select a small number of the original features that have been interpreted as an observational test for social-emotional abilities, and we could classify all subjects to distinguish those that, according to the test results, possess or do not possess social-emotional abilities. We compared the performance of two evolutionary algorithms for this task, namely ENORA and NSGA-II.

Acknowledgments

The authors acknowledge the support from the Spanish Project TIN12-39353-C04-01 and the Spanish fellowship program Ramon y Cajal RYC-2011-07821 (G. Sciavicco), and the spanish project TIN2013-4591-R (F. Jiménez and G. Sánchez). We would also like to thank the computing facilities of Extremadura Research Centre for Advanced Technologies (CETA-CIEMAT), funded by the European Regional Development Fund (ERDF); CETA-CIEMAT belongs to CIEMAT and the Government of Spain. Moreover, we would like to thank the elementary school *Colegio San Buenaventura* for collecting and providing the data, and the teaching personnel for their collaboration.

References

[1] C. Reynolds and R. Kamphaus, *Behavior Assessment for Children and Adolescent (2nd Ed.).* Circle Pines, MN: American Guidance Service, 2004.

[2] J. González, S. Fernández, E. Pérez, and P. Santamaría, *Adaptación Española del Cuestionario de Evaluación de conducta en Niños y Adolescentes (*in Spanish*)*. TEA, 2004.

[3] U. M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery: An overview," in *Advances in Knowledge Discovery and Data Mining*. AAAI, 1996, pp. 1–34.

[4] V. Kumar and S. Minz, "Feature selection: A literature review," *Smart Computing Review*, vol. 4, no. 3, pp. 211–229, 2014.

[5] R. Caruana and D. Freitag, "Greedy attribute selection," in *Proc. of the 11th International Conference on Machine Learning (ICML)*. Morgan Kaufmann, 1994, pp. 28–36.

[6] A. Arauzo-Azofra, J. Benitez, and J. Castro, "Consistency measures for feature selection," *Journal of Intelligence Information Systems*, vol. 30, no. 3, pp. 273–292, 2008.

[7] B. Blesser, T. Kuklinski, and R. Shillman, "Empirical tests for feature selection based on a psychological theory of character recognition," *Pattern Recognition*, vol. 8, no. 2, pp. 77 – 85, 1976.

[8] J. Tang and H. Liu, "Feature selection for social media data," *ACM Trans. Knowl. Discov. Data*, vol. 8, no. 4, pp. 1–27, 2014.

[9] F. Jiménez, A. Gómez-Skarmeta, G. Sánchez, and K. Deb, "An evolutionary algorithm for constrained multi-objective optimization," in *Proc. of the Congress on Evolutionary Computation (CEC)*, vol. 2. IEEE, 2002, pp. 1133–1138.

[10] F. Jiménez, G. Sánchez, and J. Juárez, "Multi-objective evolutionary algorithms for fuzzy classification in survival prediction," *Artificial Intelligence in Medicine*, vol. 60, no. 3, pp. 197–219, 2014.

[11] K. Deb, A. Pratab, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II," *IEEE Trans. on Evolutionary Computation*, vol. 6, no. 2, pp. 182 – 197, 2002.

[12] R. Sokal and P. Sneath, *Principles of Numerical Taxonomy*. W. H. Freeman, 1963.

[13] F. Wilmink and H. Uytterschaut, "Cluster analysis, history, theory and applications," in *Multivariate Statistical Methods in Physical Anthropology*, G. van Vark and W. Howells, Eds. D. Reidel Publishing Company, 1984, pp. 135–175.

[14] F. Borgen and D. Barnett, "Hierarchical cluster analysis: Comparison of three linkage measures and application to psychological data," *The Quantitative Methods for Psychology*, vol. 11, no. 1, pp. 456–468, 1987.

[15] Y. Odilia and K. Kylee, "Applying cluster analysis in counselling psychology research," *Journal of Counseling Psychology*, vol. 4, no. 34, pp. 8–21, 2015.

[16] R. Sokal and P. Sneath, *Handbook of Psychology*. Wiley, 2003.

[17] E. Shortliffe, Ed., *Computer-Based Medical Consultations: Mycin*. Elsevier, 1976.

[18] Z. Cui, "A novel medical image dynamic fuzzy classification model based onridgelet transform," *Journal of Software*, vol. 5, no. 5, pp. 456–458, 2010.

[19] I. Naresh, A. Kandel, and M. Schneider, "Feature-based fuzzy classification for interpretation of mammograms," *Fuzzy Sets and Systems*, vol. 114, no. 2, pp. 271–280, 2000.

[20] H. Liu and H. Motoda, *Feature Selection for Knowledge Discovery and Data Mining*. Kluwer, 1998.

[21] A. Marcano-Cedeno, J. Quintanilla-Dominguez, M. Cortina-Januchs, and D. Andina, "Feature selection using sequential forward selection and classification applying artificial metaplasticity neural network," in *Proc. of the 46th Annual Conference on IEEE Industrial Electronics Society (IECON)*, 2010, pp. 2845–2850.

[22] S. Cotter, K. Kreutz-Delgado, and B. Rao, "Backward sequential elimination for sparse vector subset selection," *Signal Processing*, vol. 81, no. 9, pp. 1849 – 1864, 2001.

[23] G. Nandi, "An enhanced approach to las vegas filter (LVF) feature selection algorithm," in *Proc. of the 2nd National Conference on Emerging Trends and Applications in Computer Science (NCETACS)*, 2011, pp. 1–3.

[24] H. Vafaie and K. D. Jong, "Genetic algorithms as a tool for feature selection in machine learning," in *Proc. of the 4th International Conference on Tools with Artificial Intelligence (TAI)*, 1992, pp. 200–204.

[25] S. Dreyer, "Evolutionary feature selection," Master's thesis, Institutt for datateknikk og informasjonsvitenskap, 2013.

[26] W. Siedlecki and J. Sklansky, "A note on genetic algorithms for large-scale feature selection," *Pattern Recognition Letters*, vol. 10, no. 5, pp. 335 – 347, 1989.

[27] M. ElAlami, "A filter model for feature subset selection based on genetic algorithm," *Knowledge-Based Systems*, vol. 22, no. 5, pp. 356 – 362, 2009.

[28] R. Anirudha, R. Kannan, and N. Patil, “Genetic algorithm based wrapper feature selection on hybrid prediction model for analysis of high dimensional data,” in *Proc. of the 9th International Conference on Industrial and Information Systems (ICIIS)*, 2014, pp. 1–6.

[29] H. Ishibuchi, “Multi-objective pattern and feature selection by a genetic algorithm,” in *Proc. of the Genetic and Evolutionary Computation Conference (GECCO)*, 2000, pp. 1069–1076.

[30] C. Emmanouilidis, A. Hunter, J. MacIntyre, and C. Cox, “A multi-objective genetic algorithm approach to feature selection in neural and fuzzy modeling,” *Journal of Evolutionary Optimization*, vol. 3, no. 1, pp. 1–26, 2001.

[31] A. Ekbal, S. Saha, and C. Garbe, “Feature selection using multiobjective optimization for named entity recognition,” in *Proc. of the 20th International Conference on Pattern Recognition (ICPR)*, 2010, pp. 1937–1940.

[32] Y. Jin, Ed., *Multi-Objective Machine Learning*, ser. Studies in Computational Intelligence. Springer, 2006, vol. 16.

[33] J. García-Nieto, E. Alba, L. Jourdan, and E. Talbi, “Sensitivity and specificity based multiobjective approach for feature selection: Application to cancer diagnosis,” *Information Processing Letters*, vol. 109, no. 16, pp. 887–896, 2009.

[34] A. Jara, R. Martínez, D. Vigueras, G. Sánchez, and F. Jiménez, “Attribute selection by multiobjective evolutionary computation applied to mortality from infection in severe burns patients,” in *Proc. of the International Conference on Health Informatics (HEALTHINF)*, 2011, pp. 467–471.

[35] M. Venkatadri and K. S. Rao, “A multiobjective genetic algorithm for feature selection in data mining,” *International Journal of Computer Science and Information Technologies*, vol. 1, no. 5, pp. 443–448, 2010.

[36] A. Dempster, N. Laird, and D. Rubin, “Maximum likelihood from incomplete data via the EM algorithm,” *Journal of the Royal Statistical Society*, vol. 39, no. 1, pp. 1–38, 1977.

[37] I. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems)*. Morgan Kaufmann, 2005.

[38] J. Quinlan, *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.

[39] F. Jiménez, E. Marzano, G. Sánchez, G. Sciavicco, and N. Vitacolonna, “Attribute selection via multi-objective evolutionary computation applied to multi-skill contact center data classification,” in *Proc. of the IEEE Symposium on Computational Intellgence in Big Data*, 2015, pp. 488–495.

[40] K. Deb, *Multi-Objective Optimization using Evolutionary Algorithms*. Wiley, 2001.

[41] I. Rechenberg, *Evolutionsstrategie: optimierung technischer systeme nach prinzipien der biologischen evolution*. Frommann, 1973.

[42] H. Schwefel, *Numerical Optimization of Computer Models*. Wiley, 1981.

[43] Y. Matsuyama, “Hidden markov model estimation based on alpha-EM algorithm: Discrete and continuous alpha-hmms,” in *Proc. of the 2011 International Joint Conference on Neural Networks (IJCNN)*, 2011, pp. 808–816.

[44] M. Srinivas and L. Patnaik, “Adaptive probabilities of crossover and mutation in genetic algorithms,” *IEEE Trans. on Systems, Man, and Cybernetics*, vol. 24, no. 4, pp. 656–667, 1994.

[45] “Package caret,” http://cran.r-project.org/web/packages/caret/caret.pdf, 2015.

[46] M. O’Mahony, *Sensory Evaluation of Food: Statistical Methods and Procedures*. CRC Press, 1986.

[47] J. Platt, “Sequential minimal optimization: A fast algorithm for training support vector machines,” Microsof Research, Tech. Rep., 1998, mSR-TR-98-14.

[48] E. Frank and I. Witten, “Generating accurate rule sets without global optimization,” in *Proc. of the 15th International Conference on Machine Learning*, 1998, pp. 144–151.

[49] R. Bisquerra and N. Pérez, “Las competencias emocionales (in Spanish),” *Revista Educación XXI*, vol. 10, pp. 61–82, 2007.

[50] P. Fernández and N. Ramos, *Corazones Inteligentes (in Spanish)*. Kairos, 2002.

[51] J. Payton, D. Wardlaw, P. Graczyk, M. Bloodworth, and C. T. R. Weissberg, “Social and emotional learning: A framework for promoting mental health and reducing risk behaviors in children and youth,” *Journal of School Health*, vol. 70, pp. 179–185, 2000.

[52] M. Berkowitz and M. Bier, *What Works in Character Education. A Research-Driven Guide for Educators*. Missouri University, 2005.