International Journal of Pattern Recognition
and Artificial Intelligence
Vol. 23, No. 8 (2009) 1527–1548
© World Scientific Publishing Company

DIAGNOSE EFFECTIVE EVOLUTIONARY PROTOTYPE SELECTION USING AN OVERLAPPING MEASURE*

SALVADOR GARCÍA

Department of Computer Science and Artificial Intelligence
University of Granada, Granada 18071, Spain
salvagl@decsai.ugr.es

JOSÉ-RAMÓN CANO

Department of Computer Science, University of Jaén
Higher Polytechnic Center of Linares, Alfonso X El Sabio street
Linares 23700, Spain
jrcano@ujaen.es

ESTER BERNADÓ-MANSILLA

Department of Computer Engineering, University of Ramon Llull
Barcelona 08022, Spain
esterb@salleurl.edu

FRANCISCO HERRERA

Department of Computer Science and Artificial Intelligence
University of Granada, Granada 18071, Spain
herrera@decsai.ugr.es

Evolutionary prototype selection has shown its effectiveness in the past in the prototype selection domain. It improves in most of the cases the results offered by classical prototype selection algorithms but its computational cost is expensive. In this paper, we analyze the behavior of the evolutionary prototype selection strategy, considering a complexity measure for classification problems based on overlapping. In addition, we have analyzed different k values for the nearest neighbour classifier in this domain of study to see its influence on the results of PS methods. The objective consists of predicting when the evolutionary prototype selection is effective for a particular problem, based on this overlapping measure.

Keywords: Prototype selection; evolutionary prototype selection; complexity measures; overlapping measure; data complexity.

*This work was supported by Projects TIN2005-08386-C05-01, TIN2005-08386-C05-03 and TIN2005-08386-C05-04.

1. Introduction

Prototype Selection (PS) is a classical supervised learning problem where the objective consists in, using an input data set, finding those prototypes which improve the accuracy of the nearest neighbour classifier.[28] More formally, let us assume that there is a training set T which consists of pairs (x_i, y_i), $i = 1, \ldots, n$, where x_i defines an input vector of attributes and y_i defines the corresponding class label. T contains n samples, which have m input attributes each, and they should belong to one of the C classes. Let $S \subseteq T$ be the subset of selected samples resulting from the execution of a prototype selection algorithm. The small size of the subset selected decreases the requirements of computational resources of the classification algorithm while keeping the classification performance.[1]

In the literature, another process used for reducing the number of instances can be found. This is the prototype generation, which consists of building new examples.[20,21] Many of the examples generated may not coincide with the examples belonging to the original data set, due to the fact that they are artificially generated. In some applications, this behavior is not desired, as it could be the case in some data set from the UCI Repository, such as Adult or KDD Cup'99, where information appears about real people or real connections, respectively and if new instances were generated, it could be possible that they do not correspond to valid real values. In this paper, we focus our attention on the prototype selection domain, keeping the initial characteristics of the instances unchanged.

There are many proposals of prototype selection algorithms.[16,35] These methods follow different strategies for the prototype selection problem, and offer different behaviors depending on the input data set. Evolutionary algorithms are one of the most promising heuristics.

Evolutionary Algorithms (EAs)[9,15] are general-purpose search algorithms that use principles inspired by natural genetic populations to evolve solutions to problems. The basic idea is to evolve a population of chromosomes, which represents plausible solutions to the problem, by means of a competition process. EAs have been used to solve the PS problem in Refs. 5, 22 and 33 with promising results.

The EAs offer optimal results but at the expense of high computational cost. Thus, it would be interesting to characterize their effective use in large-scale classification problems beforehand.[36] We consider their work as effective when they improve the classification capabilities of the nearest neighbors classifier. To reach this objective, we analyze the data sets characteristics prior to the prototype selection process.

In the literature, several studies have addressed the characterization of the data set by means of a set of complexity measures.[2,19] Mollineda *et al.* in Ref. 25 presented a previous work where they analyzed complexity measures like overlapping and non-parametric separability considering the Wilson's Edited Nearest Neighbor[34] and the Hart's Condensed Nearest Neighbor[17] as prototype selection algorithms.

In this study, we are interested in diagnozing when the evolutionary prototype selection is effective for a particular problem, using the overlapping measure suggested in Ref. 25. To address this, we have analyzed its behavior by means of statistical comparisons with classical prototype selection algorithms considering data sets from the UCI Repository[26] and different values of k neighbors for the prototype selection problem.

In order to do that, the paper is set out as follows. Section 2 is devoted to describe the evolutionary prototype selection strategy and the algorithm used in this study which belongs to this family. In Sec. 3, we present the complexity measure considered. Section 4 explains the experimental study and Sec. 5 deals with the results and their statistical analysis. Finally, in Sec. 6, we point out our conclusions.

2. Evolutionary Prototype Selection

EAs have been extensively used in the past both in learning and preprocessing.[5,6,11,27] EAs may be applied to the PS problem[5] because it can be considered as a search problem.

The application of EAs to PS is accomplished by tackling two important issues: the specification of the representation of the solutions and the definition of the fitness function.

- Representation: Let us assume a training data set denoted T with n instances. The search space associated with the instance selection is constituted by all the subsets of T. A chromosome consists of the sequence of n genes (one for each instance in T) with two possible states: 1 and 0, meaning that the instance is or not included in the subset selected respectively (see Fig. 1).
- Fitness function: Let S be a subset of instances coded in a chromosome that needs to be evaluated. We define the fitness function that combines two values: the classification performance (*clasper*) associated with S and the percentage of reduction (*percred*) of instances of S with respect to T:

$$\text{Fitness}(S) = \alpha \cdot \text{clasper} + (1 - \alpha) \cdot \text{percred}. \tag{1}$$

where $0 \leq \alpha \leq 1$ is the relative weight of these objectives. The k-Nearest Neighbor (k-NN) classifier is used for measuring the classification rate, *clasper*, associated with S. It denotes the percentage of objects from T correctly classified using only

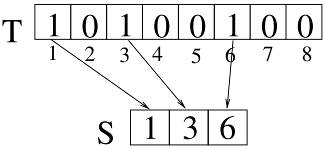

Fig. 1. Chromosome binary representation of a solution.

S to find the nearest neighbors. For each object y in T, the nearest neighbors are searched among those in the set $S\backslash\{y\}$, whereas, *percred* is defined as:

$$percred = 100 \cdot \frac{\mid T \mid - \mid S \mid}{\mid T \mid}. \tag{2}$$

The objective of the EAs is to maximize the fitness function defined, i.e. maximize the classification rate and minimize the number of instances obtained. We have used the value $\alpha = 0.5$ considering the suggestion of the authors.[5]

As EA, we have selected the CHC[10] model. This decision is based on its best competitive behaviour showed in Ref. 5. Figure 2 describes the evolutionary prototype selection process.

During each generation, Evolutionary Instance Selection CHC (EIS-CHC) method develops the following steps:

(1) It uses a parent population to generate an intermediate population of individuals, which are randomly paired and used to generate potential offspring.
(2) Then, a survival competition is held where the best chromosomes from the parent and offspring populations are selected to form the next generation.

EIS-CHC also implements a form of heterogeneous recombination using HUX, a special recombination operator. HUX exchanges half of the bits that differ between parents, where the bit position to be exchanged is randomly determined. EIS-CHC also employs a method of incest prevention. Before applying HUX to two parents, the Hamming distance between them is measured. Only those parents who differ from each other by some number of bits (mating threshold) are mated. The initial threshold is set at $L/4$, where L is the length of the chromosomes. If no offspring are inserted into the new population then the threshold is reduced by one.

No mutation is applied during the recombination phase. Instead, when the population converges or the search stops making progress (i.e. the difference threshold has dropped to zero and no new offspring are being generated which are better than

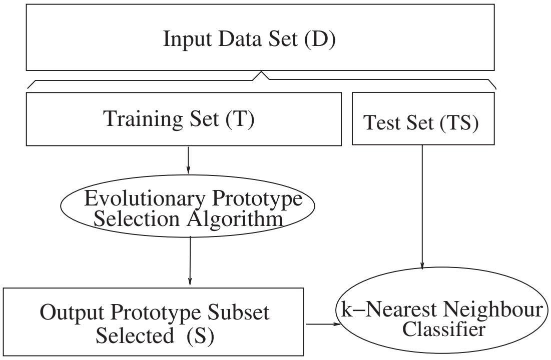
Input Data Set (D)
Training Set (T)
Test Set (TS)
Evolutionary Prototype Selection Algorithm
Output Prototype Subset Selected (S)
k–Nearest Neighbour Classifier

Fig. 2. Evolutionary prototype selection process.

any members of the parent population) the population is reinitialized to introduce new diversity to the search. The chromosome representing the best solution found along the search is used as a template to reseed the population. Reseeding of the population is accomplished by randomly changing 35% of the bits in the template chromosome to form each of the other $N - 1$ new chromosomes in the population. The search is then resumed.

The fitness function (see expression 1) combines two values: the classification rate (using k-NN) associated with S and the percentage of reduction of instances of S with respect to T.

The pseudocode of CHC appears in Algorithm 1.

3. Data Set Characterization Measure

Classification problems can be difficult for three reasons:[19]

- Certain problems are known to have nonzero Bayes error.[18] This is because some classes can be intrinsically ambiguous or due to inadequate feature measurements.
- Some problems may present complex decision boundaries so it is not possible to offer a compact description of them.[30]

```
input  : A population of chromosomes $P_a$
output: An optimized population of chromosomes $P_a$

t ← 0;
Initialize($P_a$,ConvergenceCount);
while not EndingCondition(t,$P_a$) do
    Parents ← SelectionParents($P_a$);
    Offspring ← HUX(Parents);
    Evaluate(Offspring);
    $P_n$ ← ElitistSelection(Offspring,$P_a$);
    if not modified($P_a$,$P_n$) then
        ConvergenceCount ← ConvergenceCount −1;
        if ConvergenceCount = 0 then
            $P_n$ ← Restart($P_a$);
            Initialize(ConvergenceCount);
        end
    end
    t ← t +1;
    $P_a$ ← $P_n$ ;
end
```

Algorithm 1: Pseudocode of CHC algorithm

- Sparsity induced by small sample size and high dimensionality affect the generalization of the rules.[24,29]

Real life problems are often affected by a mixture of the three previously mentioned situations.

The prediction capabilities of classifiers are strongly dependent on data complexity. This is the reason why various recent papers have introduced the use of measures to characterize the data and to relate these characteristics to the classifier performance.[30]

In Ref. 19, Ho and Basu define some complexity measures for classification problems of two classes. Singh in Ref. 32 offers a review of data complexity measures and proposes two new ones. Dong and Kothari in Ref. 8 propose a feature selection algorithm based on a complexity measure defined by Ho and Basu. Bernadó and Ho in Ref. 4 investigate the domain of competence of XCS by means of a methodology that characterizes the complexity of a classification problem by a set of geometrical descriptors. In Ref. 23, Li *et al.* analyze some omnivariate decision trees using the measure of complexity based in data density proposed by Ho and Basu. Baumgartner and Somorjai in Ref. 3 define specific measures for regularized linear classifiers, using Ho and Basu's measures as reference. Mollineda *et al.* in Ref. 25 extend some Ho and Basu's measure definitions for problems with two or more classes. They analyze these generalized measures in two classic PS algorithms and remark that Fisher's discriminant ratio is the most effective for PS. Sánchez *et al.* in Ref. 30 analyze the effect of the data complexity in the nearest neighbors classifier.

In this case, according to the conclusions of Mollineda *et al.*,[25] we have considered Fisher's discriminant ratio, which is a geometrical measure of overlapping, for studying the behavior of evolutionary prototype selection. Fisher's discriminant ratio is presented in this section.

The plain version of Fisher's discriminant ratio offered by Ho and Basu[19] computes the degree of separability of two classes according to a specific feature. It compares the difference between the class means with respect to the sum of class variances. Fisher's discriminant ratio is defined as follows:

$$f = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2} \tag{3}$$

where μ_1, μ_2, σ_1^2, σ_2^2 are the means and the variances of the two classes, respectively.

A possible generalization for C classes is proposed by Mollineda *et al.*,[25] and considers all feature dimensions. Its expression is the following:

$$F1 = \frac{\sum_{i=1}^{C} n_i \cdot \delta(m, m_i)}{\sum_{i=1}^{C} \sum_{j=1}^{n_i} \delta(x_j^i, m_i)} \tag{4}$$

where n_i denotes the number of samples in class i, δ is the metric, m is the overall mean, m_i is the mean of class i, and x_i^j represents the sample j belonging to class i. Small values of this measure indicate that classes present strong overlapping.

4. Experimental Framework

To analyze the behavior of EIS-CHC we include in the study two classical prototype selection algorithms and three advanced methods, which will be described in Sec. 4.1. In Sec. 4.2 we present the algorithms' parameters and data sets considered.

4.1. *Prototype selection algorithms*

The classical PS algorithms used in this study are: an edition algorithm (Edited Nearest Neighbor[34]) and a boundary conservative or condensation algorithm (Condensed Nearest Neighbor[17]). The advanced methods used in the comparison are: an edition method (Edition by Normalized Radial Basis Function[16]), a condensation method (Modified Selective Subset[1]) and a hybrid method, which combines edition and condensation (Decremental Reduction Optimization Procedure[35]). The use of edition schemes is motivated by the relevance of the analysis of data sets with low overlapping, where there are noisy instances inside the classes, not just in the boundaries. This is a situation where the filter PS algorithms could present an interesting behavior. The use of condensation methods is the objective of the study of the effect introduced by PS algorithms which keeps the instances situated in the boundaries, where the overlapping appears.

Their description is the following:

- Edited Nearest Neighbor (ENN).[34] The algorithm starts with $S = T$ and then each instance in S is removed if it does not agree with the majority of its k nearest neighbors. ENN filter is considered the standard noise filter and it is usually employed at the beginning of many algorithms. The pseudocode of ENN appears in Algorithm 2.
- Condensed Nearest Neighbor (CNN).[17] It begins by randomly selecting one instance belonging to each output class from T and putting them in S. Then, each instance in T is classified using only the instances in S. If an instance is misclassified, it is added to S, thus ensuring that it will be classified correctly. This

```
input  : Training set of examples T
output: Subset of training examples S

S ← T ;
foreach example $x_i$ in S do
    if $x_i$ is misclassified by its k nearest neighbours in S then
        S ← S −{$x_i$};
    end
end
return S ;
```

Algorithm 2: Pseudocode of ENN algorithm

```
input : Training set of examples T
output: Subset of training examples S
S ← $\emptyset$ ;
fail ← true;
S ← S $\bigcup\{x_{c_1}, x_{c_2}, ..., x_{c_C}\}$, where $x_{c_i}$ is any example that belongs to class $i$;
while fail = true do
    fail ← false;
    foreach example $x_i$ in T do
        if $x_i$ is misclassified by using S then
            S ← S $\bigcup\{x_i\}$;
            fail ← true;
        end
    end
end
return S ;
```

Algorithm 3: Pseudocode of CNN algorithm

process is repeated until there are no misclassified instances in T. The pseudocode of CNN appears in Algorithm 3.

- Modified Selective Algorithm (MSS).[1] Let R_i be the set of all x_i in T such that x_j is of the same class of x_i and is closer to x_i than the nearest neighbor of x_i in T of a different class than x_i. Then, MSS is defined as that subset of T containing, for every x_i in T, that element of its R_i that is the nearest to a different class than that of x_i. An efficient algorithmic representation of MSS method is depicted in Algorithm 4.
- Edition by Normalized Radial Basis Function (ENRBF).[16] It is an Edition algorithm based on the principles of Normalized Radial Basis Functions (NRBF). NRBF estimates the probability of cth class given a vector x and training set T:

$$P(c|x, T) = \sum_{i \in I^c} \bar{G}_i(x; x_i), \tag{5}$$

where $I^c = \{i : (x_i, y_i) \in T \wedge y_i = c\}$, and $\bar{G}_i(x; x_i)$ is defined by

$$\bar{G}_i(x; x_i) = \frac{G(x; x_i, \sigma)}{\sum_{j=1}^{n} G(x; x_j, \sigma)}, \tag{6}$$

and $G(x; x_i, \sigma)$ (σ is fixed) is defined by $G(x; x_i, \sigma) = e^{-\frac{||x - x_i||^2}{\sigma}}$.

ENRBF eliminates all vectors if only:

$$\exists_{c \neq y_i} \; P(y_i|x, T^i) < \alpha P(c|x, T^i). \tag{7}$$

- Decremental Reduction Optimization Procedure 3 (DROP3).[35] Its removal criterion can be restated as: *Remove x_i if at least as many of its associates in T would be classified correctly without x_i.* Each instance x_i in T continues keeping

```
input  : Training set of examples T
output: Subset of training examples S
Q ← T;
Sort the examples $\{x_j\}_{j=1}^{n}$ according to increasing values of enemy distance $D_j$;
foreach example $x_i$ in T do
    add ← false;
    foreach example $x_j$ in T do
        if $x_j \in$ Q and $d(x_i, x_j) < D_j$ then
            Q ← Q $-\{x_j\}$;
            add ← true;
        end
    end
    if add then S ← S $\bigcup\{x_i\}$;
    if Q = ∅ then return S;
end
```

Algorithm 4: Pseudocode of MSS algorithm

a list of its $k+1$ nearest neighbors in S, even after x_i is removed from S. This means that instances in S have associates that are both in and out of S, while instances that have been removed from S have no associates. DROP3 changes the order of removal of instances. It initially sorts the instances in S by the distance to their nearest enemy. Instances are then checked for removal beginning at the furthest instance from its nearest enemy. Additionally, DROP3 employs a noise filter based on ENN at the beginning of this process.

4.2. *Data sets and parameters*

The experimental study is defined in two aspects: data sets and algorithms' parameters. They are as follows:

- Data Sets: The data sets used have been collected from the UCI Repository[26] and their characteristics appear in Table 1. We consider using a ten-fold cross-validation in all data sets.
- Parameters: The parameters are chosen considering the authors' suggestions in the literature. For each one of the algorithms there are:

(a) CNN: It has not any parameter to be fixed.
(b) ENN: Minimum number of neighbours (k) = 3.
(c) MSS: It has not any parameter to be fixed.
(d) ENRBF: $\sigma = 1.0$ and $\alpha = 1.0$.
(e) DROP3: It has not any parameter to be fixed.
(f) EIS-CHC: evaluations = 10000, population = 50 and $\alpha = 0.5$.

Table 1. Data sets.

	Instances	Features	Classes
Australian	689	14	2
Balanced	625	4	3
Bupa	345	7	2
Car	1727	6	4
Cleveland	297	13	5
Contraceptive	1472	9	3
Crx	689	15	2
Dermatology	366	34	6
Ecoli	336	7	2
Glass	214	9	7
Haberman	305	3	2
Iris	150	4	3
Led7digit	500	7	10
Lymphography	148	18	4
Monks	432	6	2
New-Thyroid	215	5	3
Penbased	10992	16	10
Pima	768	8	2
Vehicle	846	18	4
Wine	178	13	3
Wisconsin	683	9	2
Satimage	6435	36	7
Thyroid	7200	21	3
Zoo	100	16	7

The algorithms were run three times for each partition in the ten-fold cross-validation. The measure F1 is calculated by averaging the F1 obtained in each training set of the ten-fold cross-validation.

5. Results and Analysis

This section contains the results and their statistical analysis, considering different values of k for the k-NN classifier to study the effect of the data complexity and the prototype selection.

We present the results obtained after the evaluation of the data sets by the prototype selection algorithms. The results for 1-NN, 3-NN and 5-NN are presented in Tables 3, 5 and 7, respectively, whose structure is the following:

- In the first column we offer the name of the data sets, ordered increasingly considering measure F1.
- The second column contains measure F1 computed for the data set in increasing order.
- The third column shows the mean test accuracy rate offered by the k-NN classifier in each data set.

- The following columns present the mean test accuracy rate and the mean reduction rate offered by CNN, ENN, MSS, ENRBF, DROP3 and EIS-CHC respectively.

In Tables 3, 5 and 7, the values in bold indicate that the test accuracy rates are equal to or higher than the ones offered by the k-NN (1-NN, 3-NN or 5-NN) using the whole data set (that is, without a previous PS process). The separation line in the tables, fixed in F1 $= 0.410$, is based on the previous works of Mollineda *et al.*[25] used as reference.

Associated to each table, we have included a statistical study based on Wilcoxon's test (see Appendix A to find its description) to analyze the behavior of the algorithms. This test allows us to establish a comparison over multiple data sets,[7,13,14] considering those delimited by the F1 measure. Tables 2, 4 and 6 show the statistical results corresponding to 1-NN, 3-NN and 5-NN, respectively. The structure of these tables is the following:

- The first column indicates the result of the test considering a level of significance $\alpha = 0.10$. With the symbol > we represent that the first algorithm outperforms the second one. The symbol = denotes that both algorithms behave equally and finally in the < case, the first algorithm is worse than the second one.
- The second column is the sum of the rankings associated to the first algorithm (see Appendix A for more details).
- The third column is the sum of the rankings related to the second algorithm.
- The fourth column shows the p-value.

In the following subsections we present the results for the case of 1-NN (Sec. 5.1), 3-NN (Sec. 5.2) and finally 5-NN (Sec. 5.3).

5.1. *Results and analysis for the 1-nearest neighbor classifier*

Tables 2 and 3 present the results of the 1-NN case.

Table 2. Wilcoxon test over 1-NN.

	1-NN with F1 < 0.410.				1-NN with F1 > 0.410.		
WILCOXON	R+	R-	p-value	WILCOXON	R+	R-	p-value
1-NN > CNN	46	9	0.059	1-NN > CNN	98	7	0.004
1-NN < ENN	5	50	0.022	1-NN = ENN	45.5	59.5	0.6
1-NN > MSS	53	2	0.009	1-NN > MSS	89.5	15.5	0.023
1-NN = ENRBF	12	33	0.214	1-NN > ENRBF	82.5	22.5	0.046
1-NN = DROP3	36	9	0.11	1-NN > DROP3	97	8	0.005
1-NN < EIS-CHC	1	54	0.007	1-NN = EIS-CHC	57	48	0.778
EIS-CHC > CNN	55	0	0.005	EIS-CHC = CNN	75	30	0.158
EIS-CHC > ENN	55	0	0.005	EIS-CHC = ENN	48	57	0.778
EIS-CHC > MSS	55	0	0.005	EIS-CHC = MSS	61	44	0.594
EIS-CHC > ENRBF	47	8	0.047	EIS-CHC > ENRBF	99	6	0.004
EIS-CHC > DROP3	55	0	0.005	EIS-CHC > DROP3	95	10	0.008

Table 3. Results considering the 1-NN classifier.

Data Set	F1	Accur. 1-NN	Accur. CNN	Red.	Accur. ENN	Red.	Accur. MSS	Red.	Accur. ENRBF	Red.	Accur. DROP3	Red.	Accur. EIS-CHC	Red.
Thyroid	0.035	0.9258	0.9007	0.8060	**0.9379**	0.0608	0.9161	0.6998	**0.9258**	0.0742	0.8581	0.9566	**0.9406**	0.9991
Lymphography	0.051	0.7387	0.7052	0.5376	**0.7547**	0.2146	**0.7436**	0.4084	**0.7609**	0.1546	**0.7674**	0.8356	**0.7938**	0.9572
Bupa	0.166	**0.6108**	0.5907	0.4180	0.5859	0.3775	0.5965	0.2055	0.5789	0.4203	0.5986	0.7005	0.6058	0.9752
Haberman	0.169	0.6697	0.6465	0.4622	**0.6990**	0.3032	0.6405	0.3678	**0.7353**	0.2647	**0.6697**	0.8003	**0.7154**	0.9855
Pima	0.217	0.7033	0.6525	0.5080	**0.7449**	0.2617	0.6797	0.3385	0.6511	0.3490	0.6901	0.8213	**0.7501**	0.9871
Contraceptive	0.224	0.4277	0.4128	0.2679	**0.4528**	0.5498	0.4196	0.1621	**0.4481**	0.5500	**0.4406**	0.7307	**0.5180**	0.9918
Cleveland	0.235	0.5314	0.5052	0.3905	**0.5576**	0.4554	0.5282	0.3099	**0.5644**	0.4404	0.4947	0.8317	**0.6173**	0.9864
Crx	0.285	0.7957	0.7913	0.6692	**0.8449**	0.2013	0.7841	0.4976	**0.8551**	0.1417	0.7841	0.8779	**0.8522**	0.9915
Australian	0.287	0.8145	0.8043	0.6441	**0.8377**	0.1514	0.8043	0.5074	**0.8609**	0.1398	0.8014	0.8847	**0.8420**	0.9918
Monks	0.365	0.7791	**0.8252**	0.6970	**0.7817**	0.0411	0.7445	0.3274	0.7619	0.2059	0.6571	0.8670	**0.9727**	0.9915
Balanced	0.455	0.7904	0.7168	0.6345	**0.8560**	0.1559	0.7825	0.3755	**0.8559**	0.1150	**0.8177**	0.8676	**0.8929**	0.9883
Dermatology	0.473	0.9535	0.9454	0.8689	**0.9591**	0.0528	0.9372	0.7450	**0.9618**	0.0404	0.9293	0.9232	**0.9617**	0.9736
Vehicle	0.506	**0.7010**	0.6655	0.5035	0.6963	0.2931	0.6868	0.3740	0.5746	0.5074	0.6099	0.7727	0.6147	0.9757
New-thyroid	0.731	**0.9723**	0.9487	0.8646	0.9494	0.0579	0.9723	0.7829	0.6981	0.3023	0.9491	0.8853	0.9541	0.9767
Glass	0.745	**0.7361**	0.6973	0.5146	0.6919	0.3214	0.7206	0.3905	0.3956	0.6774	0.6571	0.7415	0.6227	0.9507
Ecoli	0.888	0.8070	0.7532	0.5989	**0.8248**	0.1958	0.7800	0.4726	0.4256	0.5744	0.7263	0.8423	0.8039	0.9653
Zoo	0.949	**0.9281**	0.8289	0.7579	0.9114	0.0815	0.8467	0.7834	0.9281	0.0571	0.9264	0.8351	0.9142	0.9111
Car	1.022	0.8565	**0.8825**	0.7658	0.8513	0.0775	0.8455	0.5489	0.7002	0.2998	0.6221	0.8834	**0.8802**	0.9897
Penbased	1.161	**0.9935**	0.9853	0.9555	0.9927	0.0065	0.9913	0.9051	0.8834	0.1902	0.9431	0.9719	0.9560	0.9871
Led7digit	1.344	0.4020	0.3740	0.2727	**0.4920**	0.5660	**0.4020**	0.1629	**0.4580**	0.2609	**0.4060**	0.8278	**0.6580**	0.9649
Wisconsin	1.354	0.9557	0.9342	0.8965	**0.9657**	0.0529	0.9485	0.8210	0.9414	0.0769	0.8913	0.9747	**0.9642**	0.9941
Satimage	1.476	**0.9058**	0.8835	0.8041	0.9013	0.0890	0.9009	0.6841	0.7302	0.4395	0.8308	0.9101	0.8710	0.9950
Wine	1.820	0.9552	0.9157	0.8308	**0.9552**	0.0337	**0.9605**	0.7247	**0.9663**	0.0543	0.9105	0.9101	0.9438	0.9732
Iris	2.670	0.9333	0.9200	0.8519	**0.9533**	0.0474	**0.9467**	0.7844	**0.9400**	0.1267	0.9267	0.9230	**0.9733**	0.9674

The analysis of Tables 2 and 3 is the following:

- *F1 low [0,0.410] which represents strong overlapping*: The evolutionary algorithm (EIS-CHC) outperforms 1-NN when F1 is low. EIS-CHC presents the best accuracy rates among all the PS algorithms in most of the data sets with the strongest overlapping. Wilcoxon's test supports this observation (Table 2).
- *F1 high [0.410,...], being small overlapping*: There is not any improvement of PS algorithms with respect to Without PS, as statistical results indicate. The benefit of the use of the PS algorithms in these kind of data sets using the 1-NN is the reduction of the size of the data set. Only ENN and EIS-CHC obtain the same performance as not using PS. The comparison between EIS-CHC and the rest of the models indicates that the accuracy of EIS-CHC is always better than or equal to that of the method compared (Table 2).

Considering the results that CNN and MSS present in Table 3 we must point out that the PS algorithms which keep boundary instances (condensation methods) notably affect the classification capabilities of the 1-NN classifier, independently of the overlapping of the data set. DROP3 obtains a performance similar to that of not using PS, due to the fact that it integrates a noise filter in its definition.

Paying attention to the relation between F1 and the behavior of EIS-CHC, we can point out that the use of this measure can help us to decide when the use of EIS-CHC improves the accuracy rates of 1-NN classifier in a concrete data set, previously to its execution.

5.2. *Results and analysis for the 3-nearest neighbor classifier*

Tables 4 and 5 present the results of the PS methods with the 3-NN classifier.

The analysis of Tables 4 and 5 is the following:

- *F1 low [0,0.410]* (strong overlapping): Similarly to the case with the 1-NN, EIS-CHC outperforms the Without PS when F1 is low. EIS-CHC presents the best accuracy rates among all the PS algorithms in all data sets with the strongest overlapping. Wilcoxon's test in Table 4 confirms this affirmation.
- *F1 high [0.410,...]* (small overlapping): The situation is similar to the previous case. There is no improvement of PS algorithms with respect to 3-NN, as the statistical results indicate (see Table 4). Only ENN and EIS-CHC obtain the same performance as not using PS. The comparison between EIS-CHC and the rest of the models indicates that the accuracy of EIS-CHC is always better than or equal to that of the method compared.

Note that when $k = 3$, the nearest neighbor classifier is more robust in the presence of noise than the 1-NN classifier. Due to this fact, the ENN and ENRBF filters behave similarly to the 3-NN when F1 is lower than 0.410, according to Wilcoxon's test. The same effect occurs in DROP3. However, a PS process by

Table 4. Wilcoxon test over 3-NN.

	3-NN with F1 < 0.410.				3-NN with F1 > 0.410.		
WILCOXON	R+	R-	p-value	WILCOXON	R+	R-	p-value
3-NN > CNN	53	2	0.009	3-NN > CNN	104.5	0.5	0.001
3-NN = ENN	15.5	39.5	0.26	3-NN = ENN	63.5	41.5	0.507
3-NN = MSS	43	12	0.114	3-NN > MSS	82.5	22.5	0.064
3-NN = ENRBF	31.5	23.3	0.594	3-NN > ENRBF	91	14	0.016
3-NN > DROP3	51	4	0.017	3-NN > DROP3	99	6	0.004
3-NN < EIS-CHC	6	49	0.028	3-NN = EIS-CHC	75	30	0.158
EIS-CHC > CNN	55	0	0.005	EIS-CHC = CNN	54	51	0.925
EIS-CHC > ENN	50.5	4.5	0.021	EIS-CHC = ENN	30	75	0.158
EIS-CHC > MSS	51	4	0.017	EIS-CHC = MSS	46	59	0.683
EIS-CHC > ENRBF	47	8	0.047	EIS-CHC > ENRBF	90	15	0.019
EIS-CHC > DROP3	55	0	0.005	EIS-CHC > DROP3	104	1	0.001

EIS-CHC prior to the 3-NN classifier improves the accuracy of the classifier without using PS and also achieves a high reduction of the subset selected.

5.3. *Results and analysis for the 5-nearest neighbor classifier*

Tables 6 and 7 present the results of the PS methods with the 5-NN classifier.

The analysis of Tables 6 and 7 is the following:

- *F1 low [0,0.410]* (strong overlapping): EIS-CHC outperforms the Without PS when F1 is low. EIS-CHC presents the best accuracy rates among all the PS algorithms in most of the data sets with the strongest overlapping (Table 7). Considering Wilcoxon's test in Table 6, only EIS-CHC improves the classification capabilities of 5-NN which reflects the proper election of the most representative instances in the presence of overlapping.
- *F1 high [0.410,...]* (small overlapping): The situation is similar to the previous case. There is no improvement of PS algorithms with respect to 5-NN, as the statistical results indicate (see Table 6). Only ENN and EIS-CHC obtain the same performance as not using PS. The comparison between EIS-CHC and the rest of models indicates that the accuracy of EIS-CHC is always better than or equal to that of the method compared.

In this case, ENN and ENRBF obtain a result similar to the previous subsection (3-NN case), where F1 is low, but again EIS-CHC offers a significant improvement in accuracy with respect to the use of the nearest neighbors classifier without using PS.

5.4. *Summary of the analysis*

Considering the previous results and analysis we can present as summary the following comments:

- Independently of the k value selected for the nearest neighbors classifier, when the overlapping of the initial data set is strong (it presents low values of F1)

Table 5. Results considering the 3-NN classifier.

		Accur.	Accur.		Accur.		Accur.		Accur.		Accur.		Accur.	
Data Set	F1	3-NN	CNN	Red.	ENN	Red.	MSS	Red.	ENRBF	Red.	DROP3	Red.	EIS-CHC	Red.
Thyroid	0.035	0.9236	0.9083	0.7718	**0.9250**	0.0759	**0.9360**	0.6998	**0.9258**	0.0742	0.8056	0.9812	**0.9250**	0.9983
Lymphography	0.051	0.7739	**0.7826**	0.5580	0.7530	0.2146	0.7589	0.4084	0.7530	0.1546	0.7053	0.8679	**0.8034**	0.9535
Bupa	0.166	0.6066	0.5845	0.3649	**0.6174**	0.3775	**0.6112**	0.2055	0.5789	0.4203	**0.6315**	0.7649	**0.6524**	0.9755
Haberman	0.169	0.7058	0.6537	0.4281	**0.7125**	0.3032	0.6959	0.3678	**0.7353**	0.2647	0.6500	0.8885	**0.7219**	0.9862
Pima	0.217	0.7306	0.6654	0.5049	**0.7384**	0.2617	0.7176	0.3385	0.6511	0.3490	0.7176	0.8562	**0.7684**	0.9860
Contraceptive	0.224	0.4495	0.4447	0.2475	**0.4583**	0.5498	**0.4528**	0.1621	**0.4522**	0.5500	**0.4542**	0.7442	**0.4875**	0.9911
Cleveland	0.235	0.5444	0.5247	0.3935	**0.5447**	0.4448	0.5346	0.3099	**0.5677**	0.4404	0.4688	0.8775	**0.5643**	0.9802
Crx	0.285	0.8420	0.8203	0.6531	**0.8449**	0.1560	0.8304	0.4976	**0.8522**	0.1417	0.7377	0.9274	**0.8536**	0.9905
Australian	0.287	0.8478	0.8203	0.6581	**0.8478**	0.1514	0.8348	0.5074	**0.8478**	0.1398	0.7783	0.9293	**0.8681**	0.9897
Monks	0.365	0.9629	0.8658	0.6556	**0.9567**	0.0411	0.9613	0.3274	0.8143	0.2059	0.6957	0.8678	**0.9376**	0.9830
Balanced	0.455	0.8337	0.7472	0.6240	**0.8768**	0.1559	0.8304	0.3755	**0.8768**	0.1150	0.8193	0.9090	**0.9009**	0.9860
Dermatology	0.473	**0.9700**	0.9511	0.8607	0.9619	0.0310	0.9457	0.7450	0.9646	0.0404	0.8987	0.9262	0.9429	0.9505
Vehicle	0.506	**0.7175**	0.6619	0.4768	0.6927	0.2931	0.6809	0.3740	0.5569	0.5074	0.5545	0.8164	0.6087	0.9634
New-thyroid	0.731	0.9537	0.9485	0.8476	0.9307	0.0579	0.9394	0.7829	0.6981	0.3023	0.8799	0.9189	**0.9584**	0.9592
Glass	0.745	**0.7011**	0.6493	0.4762	0.6616	0.3214	0.6650	0.3905	0.3842	0.6774	0.5764	0.8105	0.6267	0.9465
Ecoli	0.888	0.8067	0.7650	0.5913	**0.8126**	0.1958	0.7767	0.4726	0.4256	0.5744	0.6636	0.8714	0.7534	0.9583
Zoo	0.949	**0.9281**	0.8328	0.6954	0.9114	0.0815	0.7811	0.7834	0.9114	0.0571	0.8436	0.7859	0.9006	0.8813
Car	1.022	**0.9231**	0.9010	0.7662	0.8930	0.0775	0.9173	0.5489	0.7002	0.2998	0.6887	0.8865	0.8409	0.9853
Penbased	1.161	0.9718	0.9536	0.8464	0.9618	0.0287	**0.9914**	0.9051	0.8799	0.1902	0.8765	0.9783	0.8700	0.9568
Led7digit	1.344	0.4520	0.4040	0.2829	**0.5460**	0.5660	0.4320	0.1629	0.4300	0.2609	**0.5320**	0.8331	**0.6900**	0.9509
Wisconsin	1.354	0.9600	0.9542	0.8894	**0.9657**	0.0337	**0.9600**	0.8210	0.9471	0.0769	0.9099	0.9777	**0.9685**	0.9908
Satimage	1.476	0.8662	0.8444	0.7171	0.8646	0.1322	**0.9061**	0.6841	0.7316	0.4395	0.7706	0.9367	0.8164	0.9666
Wine	1.820	0.9549	0.9327	0.8514	**0.9549**	0.0337	**0.9552**	0.7247	**0.9719**	0.0543	0.9154	0.9207	0.9438	0.9457
Iris	2.670	0.9400	**0.9400**	0.8415	**0.9533**	0.0474	0.9333	0.7844	**0.9467**	0.1267	0.8467	0.9326	**0.9600**	0.9333

Table 6. Wilcoxon test over 5-NN.

	5-NN with F1 < 0.410.				5-NN with F1 > 0.410.		
WILCOXON	R+	R-	p-value	WILCOXON	R+	R-	p-value
5-NN > CNN	55	0	0.005	5-NN > CNN	102.5	2.5	0.002
5-NN = ENN	25	30	0.799	5-NN = ENN	76.5	28.5	0.158
5-NN > MSS	40	5	0.038	5-NN > MSS	87	18	0.03
5-NN = ENRBF	41	14	0.169	5-NN > ENRBF	93	12	0.011
5-NN > DROP3	40	5	0.038	5-NN > DROP3	100	5	0.003
5-NN < EIS-CHC	9	46	0.059	5-NN = EIS-CHC	75.5	29.5	0.136
EIS-CHC > CNN	55	0	0.005	EIS-CHC = CNN	50.5	54.5	0.9
EIS-CHC > ENN	52.5	2.5	0.011	EIS-CHC = ENN	26.5	78.5	0.116
EIS-CHC > MSS	49	6	0.028	EIS-CHC = MSS	56	49	0.826
EIS-CHC > ENRBF	52.5	2.5	0.011	EIS-CHC > ENRBF	86	19	0.035
EIS-CHC > DROP3	55	0	0.005	EIS-CHC > DROP3	104	1	0.001

EIS-CHC is a very effective PS algorithm to improve the accuracy rates of the nearest neighbors classifier.

- When the overlapping of the data set is low, the statistical test has shown that the PS algorithms are not capable of improving the accuracy of the k-NN without using PS. The benefits of their use is that they keep the accuracy capabilities of the nearest neighbors classifier, reducing the initial data set size.
- Considering the results that CNN and MSS present, we must point out that the PS algorithms which keep boundary instances (condensation methods) notably affect the classification capabilities of the k-NN classifier, independently of the overlapping of the data set and the value of k.
- In the classical algorithms, the best behavior corresponds to ENN. The filter process that ENN introduces outperforms in some cases the classification capabilities of the k-NN, but the election of the most representative prototypes that EIS-CHC develops seems to be the most effective strategy. Nevertheless, ENN in combination with k-NN obtains similar results to k-NN when $k \geq 3$, given that the nearest neighbors classifier is more robust in the presence of noise.
- In the most advanced algorithms, the behavior coincides in most of the cases with the equivalent in the classic algorithms; MSS behaves very similarly to CNN and ENRBF to ENN. DROP3, as a hybrid model alike EIS-CHC, obtains an intermediate behavior between condensation and edition methods, because it performs adequately when strong overlapping is presented when considering 1-NN. Nevertheless, EIS-CHC always outperforms DROP3 in any case.

Paying attention to the relation between F1 and the behavior of EIS-CHC, we can point out that the use of this measure can help us decide when the use of EIS-CHC improves the accuracy rates of k-NN classifier in a concrete data set, previously to its execution.

With these results in mind, we could analyze the F1 measure in a new data set and if it is small (F1 between [0,0.410)), we can use EIS-CHC as PS method to

Table 7. Results considering the 5-NN classifier.

		Accur.	Accur.		Accur.		Accur.		Accur.		Accur.		Accur.	
Data Set	F1	5-NN	CNN	Red.	ENN	Red.	MSS	Red.	ENRBF	Red.	DROP3	Red.	EIS-CHC	Red.
Thyroid	0.035	0.9292	0.9056	0.7810	0.9250	0.0716	**0.9396**	0.6998	0.9258	0.0742	0.7901	0.9842	0.9250	0.9978
Lymphography	0.051	0.7944	0.7931	0.5736	**0.7796**	0.2079	0.7922	0.4084	0.7801	0.1546	0.6983	0.8904	**0.8423**	0.9369
Bupa	0.166	0.6131	0.6101	0.3304	**0.6137**	0.3910	0.6045	0.2055	0.5789	0.4203	**0.6137**	0.8039	**0.6464**	0.9710
Haberman	0.169	0.6695	0.6402	0.4230	**0.7288**	0.3061	0.6594	0.3678	**0.7353**	0.2647	**0.7019**	0.9346	**0.7353**	0.9960
Pima	0.217	0.7306	0.7044	0.4899	**0.7396**	0.2626	**0.7306**	0.3385	0.6511	0.3490	0.7187	0.8866	**0.7671**	0.9854
Contraceptive	0.224	0.4685	0.4542	0.2383	**0.4725**	0.5374	0.4562	0.1621	0.4535	0.5500	**0.4685**	0.7690	**0.4820**	0.9909
Cleveland	0.235	0.5545	0.5382	0.3748	**0.5676**	0.4488	0.5446	0.3099	**0.5711**	0.4404	0.5222	0.8969	**0.6075**	0.9725
Crx	0.285	0.8551	0.8232	0.6659	0.8507	0.1433	0.8435	0.4976	0.8536	0.1417	0.7014	0.9406	**0.8594**	0.9874
Australian	0.287	0.8478	0.8159	0.6591	**0.8609**	0.1588	0.8304	0.5074	0.8435	0.1398	0.7188	0.9499	**0.8580**	0.9865
Monks	0.365	**0.9475**	0.8523	0.6168	0.8855	0.0432	0.9263	0.3274	0.8013	0.2059	0.7490	0.8382	0.8959	0.9784
Balanced	0.455	0.8624	0.8080	0.6574	**0.8928**	0.1351	0.8496	0.3755	**0.8831**	0.1150	0.8013	0.9234	**0.8879**	0.9838
Dermatology	0.473	**0.9646**	0.9481	0.8443	0.9592	0.0316	0.9318	0.7450	0.9619	0.0404	0.8772	0.9250	0.9426	0.9375
Vehicle	0.506	**0.7175**	0.6903	0.4641	0.6822	0.2871	0.6844	0.3740	0.5592	0.5074	0.5497	0.8240	0.6123	0.9567
New-thyroid	0.731	0.9398	**0.9400**	0.8439	0.9119	0.0646	0.9165	0.7829	0.6981	0.3023	0.9119	0.9173	**0.9680**	0.9385
Glass	0.745	**0.6685**	0.6531	0.4429	0.6652	0.3453	0.6067	0.3905	0.3609	0.6774	0.5813	0.8172	0.6331	0.9216
Ecoli	0.888	**0.8127**	0.7799	0.6035	0.8065	0.1812	0.7889	0.4726	0.4256	0.5744	0.6620	0.8810	0.7351	0.9527
Zoo	0.949	**0.9364**	0.8097	0.6044	0.8964	0.0708	0.7536	0.7834	0.9197	0.0571	0.7125	0.7506	0.8717	0.8470
Car	1.022	0.9520	0.9323	0.7748	0.9016	0.0475	0.9196	0.5489	0.7002	0.2998	0.7066	0.8749	0.8368	0.9819
Penbased	1.161	0.9618	0.9455	0.8195	0.9482	0.0376	**0.9864**	0.9051	0.8739	0.1902	0.8551	0.9774	0.8500	0.9432
Led7digit	1.344	0.4140	0.3860	0.2878	**0.5520**	0.5860	0.3940	0.1629	**0.4180**	0.2609	**0.4820**	0.8347	**0.6500**	0.9416
Wisconsin	1.354	0.9657	0.9500	0.8886	**0.9671**	0.0296	**0.9686**	0.8210	0.9485	0.0769	0.8624	0.9777	**0.9657**	0.9876
Satimage	1.476	0.8740	0.8412	0.7099	0.8708	0.1317	**0.9050**	0.6841	0.7329	0.4395	0.7416	0.9473	0.8289	0.9603
Wine	1.820	0.9605	**0.9605**	0.8421	**0.9605**	0.0418	**0.9663**	0.7247	**0.9719**	0.0543	0.9157	0.8976	**0.9660**	0.9295
Iris	2.670	**0.9600**	0.9533	0.8356	**0.9600**	0.0430	0.6267	0.7844	0.9400	0.1267	0.9133	0.9119	**0.9600**	0.9148

improve the accuracy rate of the k-NN classifier. When F1 is greater than 0.410, EIS-CHC offers interesting behavior, with accuracy equivalent to the obtained without reduction as Wilcoxon's test indicates, but with reduction rates larger than 90% in most of the data sets.

6. Concluding Remarks

This paper addresses the analysis of the evolutionary prototype selection considering a complexity data set measure based on overlapping, with the objective of predicting when the evolutionary prototype selection is effective for a particular problem.

An experimental study has been carried out using data sets from different domains and comparing the results with classical PS algorithms, having the F1 measure as reference. To extend the analysis of the k-NN classifier we have considered different values of k. The main conclusions reached are the following:

- EIS-CHC presents the best accuracy rate when the input data set has strong overlapping, even improving condensation algorithms (CNN and MSS), edition schemes (ENN and ENRBF) and hybrid methods, such as DROP3.
- EIS-CHC improves the classification accuracy of k-NN when the data sets have strong overlapping, independently of the k value, and obtains a high reduction rate of the data. However, ENN, ENRBF and DROP3 algorithms are not able to improve the accuracy rate of k-NN when $k \geq 3$.
- In the case of data sets with low overlapping, the results of the PS algorithms are not conclusive so none of them can be suggested considering accuracy rates. Therefore, their use is recommended to keep the accuracy capabilities by reducing the initial data set size.
- Condensation algorithms, which keep the boundaries (CNN and MSS), have normally shown negative effects on the accuracy of the k-NN classifier.

As we have indicated in the analysis section, the use of this measure can help us to evaluate a data set previously to the evolutionary PS process and decide if it is adequate or not to improve the classification capabilities of the k-NN classifier.

The results show that when F1 is low (strong overlapping), the best accuracy rates appear using EIS-CHC, while when F1 is high (low overlapping), the PS algorithms do not guarantee an accuracy improvement.

As future works, the analysis of the effect of data complexity on evolutionary instance selection for training set selection considering other well-known classification algorithms will be studied. Another interesting research line is the measurement of data complexity on imbalanced data sets when we can perform evolutionary under-sampling.[12]

Appendix A. Wilcoxon Signed Rank Test

Wilcoxon's test is used for answering this question: do two samples represent two different populations? It is a nonparametric procedure employed in a hypothesis testing situation involving a design with two samples. It is the analogous of the paired t-test in nonparametrical statistical procedures; therefore, it is a pairwise test that aims to detect significant differences between the behavior of two algorithms.

The null hypothesis for Wilcoxon's test is $H_0 : \theta_D = 0$; in the underlying populations represented by the two samples of results, the median of the difference scores equals zero. The alternative hypothesis is $H_1 : \theta_D \neq 0$, but $H_1 : \theta_D > 0$ or $H_1 : \theta_D < 0$ can also be used as directional hypothesis.

In the following, we describe the tests computations. Let d_i be the difference between the performance scores of the two algorithms on ith out of N data sets. The differences are ranked according to their absolute values; average ranks are assigned in case of ties. Let R^+ be the sum of ranks for the data sets on which the second algorithm outperformed the first, and R^- the sum of ranks for the opposite. Ranks of $d_i = 0$ are split evenly among the sums; if there is an odd number of them, one is ignored:

$$R^+ = \sum_{d_i>0} \mathrm{rank}(d_i) + \frac{1}{2}\sum_{d_i=0} \mathrm{rank}(d_i)$$

$$R^- = \sum_{d_i<0} \mathrm{rank}(d_i) + \frac{1}{2}\sum_{d_i=0} \mathrm{rank}(d_i)$$

Let T be the smallest of the sums, $T = \min(R^+, R^-)$. If T is less than or equal to the value of the distribution of Wilcoxon for N degrees of freedom (Table B.12 in Ref. 37), the null hypothesis of equality of means is rejected.

Obtaining the p-value associated to a comparison is performed by means of the normal approximation for the Wilcoxon T statistic (Sec. VI, Test 18 in Ref. 31). Furthermore, the computation of the p-value for this test is usually included in well-known statistical software packages (SPSS, SAS, R, etc.).

References

1. R. Barandela, F. J. Ferri and J. S. Sánchez, Decision boundary preserving prototype selection for nearest neighbor classification, *Int. J. Patt. Recogn. Artif. Intell.* **19**(6) (2005) 787–806.
2. M. Basu and T. K. Ho, *Data Complexity in Pattern Recognition* (Springer, 2006).
3. R. Baumgartner and R. L. Somorjai, Data complexity assessment in undersampled classification of high-dimensional biomedical data, *Patt. Recogn. Lett.* **27**(12) (2006) 1383–1389.
4. E. Bernadó-Mansilla and T. K. Ho, Domain of competence of XCS classifier system in complexity measurement space, *IEEE Trans. Evolut. Comput.* **9**(1) (2005) 82–104.
5. J.-R. Cano, F. Herrera and M. Lozano, Using evolutionary computation as instance selection for data reduction in KDD: An experimental study, *IEEE Trans. Evolut. Comput.* **7**(6) (2003) 561–575.
6. L. P. Corella, C. De Stefano and F. Fontanella, Evolutionary prototyping for handwriting recognition, *Int. J. Patt. Recogn. Artif. Intell.* **21**(1) (2007) 157–178.
7. J. Demsar, Statistical comparison of classifiers over multiple data sets, *J. Mach. Learn. Res.* **7** (2006) 1–30.

8. M. Dong and R. Kothari, Feature subset selection using a new definition of classifiability, *Patt. Recogn. Lett.* **24**(9–10) (2003) 1215–1225.
9. A. E. Eiben and J. E. Smith, *Introduction to Evolutionary Computation* (Springer-Verlag, 2003).
10. L. J. Eshelman, The CHC adaptive search algorithm: How to have safe search when engaging in nontraditional genetic recombination, in *Foundations of Genetic Algorithms 1*, ed. G. J. E. Rawlins (Morgan Kauffman, 1991), pp. 265–283.
11. C. Gagné and M. Parizeau, Co-evolution of nearest neighbor classifiers, *Int. J. Patt. Recogn. Artif. Intell.* **21**(5) (2007) 912–946.
12. S. García and F. Herrera, Evolutionary under-sampling for classification with imbalanced data sets: Proposals and taxonomy, *Evolut. Comput.* **17**(3) (2008) 275–306.
13. S. García and F. Herrera, An extension on statistical comparisons of classifiers over multiple data sets for all pairwise comparisons, *J. Mach. Learn. Res.* **9** (2008) 2677–2694.
14. S. García, A. Fernández, J. Luengo and F. Herrera, A study of statistical techniques and performance measures for genetics-based machine learning: Accuracy and interpretability, *Soft Comput.* **13**(10) (2009) 959–977.
15. D. E. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning* (Addison-Wesley, 1989).
16. M. Grochowski and N. Jankowski, Comparison of instance selection algorithms II. Results and comments, *Proc. 7th Int. Conf. Artificial Intelligence and Soft Computing*, Lecture Notes in Computer Science, Vol. 3070 (2004), pp. 580–585.
17. P. E. Hart, The condensed nearest neighbor rule, *IEEE Trans. Inform. Th.* **14** (1968) 515–516.
18. T. K. Ho and H. S. Baird, Large-scale simulation studies in image pattern recognition, *IEEE Trans. Patt. Anal. Mach. Intell.* **19**(10) (1997) 1067–1079.
19. T. K. Ho and M. Basu, Complexity measures of supervised classification problems, *IEEE Trans. Patt. Anal. Mach. Intell.* **24**(3) (2002) 289–300.
20. S. W. Kim and B. J. Oommen, Enhancing prototype reduction schemes with LVQ3-type algorithms, *Patt. Recogn.* **36** (2004) 1083–1093.
21. S. W. Kim and B. J. Oommen, On using prototype reduction schemes to optimize kernel-based nonlinear subspace methods, *Patt. Recogn.* **37** (2004) 227–239.
22. L. Kuncheva, Editing for the k-nearest neighbors rule by a genetic algorithm, *Patt. Recogn. Lett.* **16** (1995) 809–814.
23. Y.-H. Li, M. Dong and R. Kothari, Classifiability-based omnivariate decision trees, *IEEE Trans. Neural Networks* **16**(6) (2005) 1547–1560.
24. M. Liwicki and H. Bunke, Handwriting recognition of whiteboard notes — studying the influence of training set size and type, *Int. J. Patt. Recogn. Artif. Intell.* **21**(1) (2007) 83–98.
25. R. A. Mollineda, J. S. Sánchez and J. M. Sotoca, Data characterization for effective prototype selection, *Proc. IbPRIA 2005*, Lecture Notes in Computer Science, Vol. 3523 (2005), pp. 27–34.
26. A. Asuncion and D. J. Newman, UCI Machine Learning Repository [http://www.ics.uci.edu/∼mlearn/MLRepository.html]. Irvine, CA: University of California, Schools of Information and Computer Science, 2007.
27. I. S. Oh, J. S. Lee and B. R. Moon, Hybrid genetic algorithms for feature selection, *IEEE Trans. Patt. Anal. Mach. Intell.* **26**(11) (2004) 1424–1437.
28. A. N. Papadopoulos and Y. Manolopoulos, *Nearest Neighbor Search: A Database Perspective* (Springer-Verlag, 2004).

29. X. Qiu and L. Wu, Nearest neighbor discriminant analysis, *Int. J. Patt. Recogn. Artif. Intell.* **20**(8) (2006) 1245–1259.
30. J. S. Sánchez, R. A. Mollineda and J. M. Sotoca, An analysis of how training data complexity affects the nearest neighbors classifiers, *Patt. Anal. Appl.* **10** (2007) 189–201.
31. D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*, Second Edition (Chapman and Hall, 2003).
32. S. Singh, Multiresolution estimates of classification complexity, *IEEE Trans. Patt. Anal. Mach. Intell.* **25**(12) (2003) 1534–1539.
33. H. Shinn-Ying, L. Chia-Cheng and L. Soundy, Design of an optimal nearest neighbor classifier using an intelligent genetic algorithm, *Patt. Recogn. Lett.* **23**(13) (2002) 1495–1503.
34. D. L. Wilson, Asymptotic properties of nearest neighbor rules using edited data, *IEEE Trans. Syst. Man Cybern.* **2**(3) (1972) 408–421.
35. D. R. Wilson and T. R. Martinez, Reduction techniques for instance-based learning algorithms, *Mach. Learn.* **38** (2000) 257–268.
36. B. Yang, X. Su and Y. Wang, Distributed learning based on chips for classification with large-scale data set, *Int. J. Patt. Recogn. Artif. Intell.* **21**(5) (2007) 899–920.
37. J. H. Zar, *Biostatistical Analysis*, 4th Edition (Pearson, 1999).

Salvador García received the M.Sc. and Ph.D. degrees in computer science from the University of Granada, Spain, in 2004 and 2008, respectively. He is currently an Assistant Professor in the Department of Computer Science, University of Jaén, Spain.

His research interests include data mining, data reduction, data complexity, imbalanced learning, statistical inference and evolutionary algorithms.

José-Ramón Cano received the M.Sc. and Ph.D. degrees in computer science from the University of Granada, Spain, in 1999 and 2004, respectively. He is currently an Associate Professor in the Department of Computer Science, University of Jan, Spain.

His research interests include data mining, data reduction, data complexity, interpretability-accuracy trade-off, and evolutionary algorithms.

Ester Bernadó-Mansilla received the B.Sc. degree in telecommunications engineering in 1992, the M.Sc. degree in electronic engineering in 1995, and the Ph.D. degree in computer science in 2002, from Enginyeria i Arquitectura La Salle, Ramon Llull University, Barcelona. She is currently an associate professor in the Computer Engineering Department of Enginyeria i Arquitectura La Salle, Ramon Llull University, Barcelona, Spain. She has co-edited two books on genetic-based machine learning. She serves as associate editor of *Pattern Recognition Letters*.

Her research interests include machine learning, pattern recognition, data mining, genetic algorithms, and genetic-based machine learning.

Francisco Herrera received the M.Sc. degree in mathematics in 1988 and the Ph.D. degree in mathematics in 1991, both from the University of Granada, Spain. He is currently a Professor in the Department of Computer Science and Artificial Intelligence at the University of Granada. He has published more than 150 papers in international journals. He is coauthor of the book *Genetic Fuzzy Systems: Evolutionary Tuning and Learning of Fuzzy Knowledge Bases* (World Scientific, 2001). As editorial activities, he has co-edited five international books and co-edited twenty special issues in international journals on different Soft Computing topics. He is an associated editor of the journals: *IEEE Transactions on Fuzzy Systems*, *Information Sciences*, *Mathware and Soft Computing*, *Advances in Fuzzy Systems*, *Advances in Computational Sciences and Technology*, and *International Journal of Applied Metaheuristics Computing*. He currently serves as area editor of the *Journal on Soft Computing* (area of genetic algorithms and genetic fuzzy systems), and he serves as member of several journal editorial boards, among others: *Fuzzy Sets and Systems*, *Applied Intelligence*, *Knowledge and Information Systems*, *Information Fusion*, *Evolutionary Intelligence*, *International Journal of Hybrid Intelligent Systems*, *Memetic Computation*.

His current research interests include computing with words and decision making, data mining, data preparation, instance selection, fuzzy rule based systems, genetic fuzzy systems, knowledge extraction based on evolutionary algorithms, memetic algorithms and genetic algorithms.