Soft Comput (2010) 14:713–728
DOI 10.1007/s00500-009-0460-y

ORIGINAL PAPER

Multi-objective genetic fuzzy classifiers for imbalanced and cost-sensitive datasets

Pietro Ducange · Beatrice Lazzerini ·
Francesco Marcelloni

Abstract We exploit an evolutionary three-objective optimization algorithm to produce a Pareto front approximation composed of fuzzy rule-based classifiers (FRBCs) with different trade-offs between accuracy (expressed in terms of sensitivity and specificity) and complexity (computed as sum of the conditions in the antecedents of the classifier rules). Then, we use the ROC convex hull method to select the potentially optimal classifiers in the projection of the Pareto front approximation onto the ROC plane. Our method was tested on 13 highly imbalanced datasets and compared with 2 two-objective evolutionary approaches and one heuristic approach to FRBC generation, and with three well-known classifiers. We show by the Wilcoxon signed-rank test that our three-objective optimization approach outperforms all the other techniques, except for one classifier, in terms of the area under the ROC convex hull, an accuracy measure used to globally compare different classification approaches. Further, all the FRBCs in the ROC convex hull are characterized by a low value of complexity. Finally, we discuss how, the misclassification costs and the class distributions are fixed, we can select the most suitable classifier for the specific application. We show that the FRBC selected from the convex hull produced by our three-objective optimization approach achieves the lowest classification cost among the techniques used as comparison in two specific medical applications.

Keywords Genetic fuzzy rule-based classifiers · Multi-objective evolutionary algorithms · Imbalanced datasets · ROC curves · Convex hull method

1 Introduction

Multi-objective genetic fuzzy modeling uses genetic/evolutionary algorithms to design fuzzy rule-based systems (FRBSs) with a good trade-off between accuracy and interpretability (Herrera 2008; Ishibuchi 2007). As discussed in the specialized literature (Casillas et al. 2003a; Cordon et al. 2001; Herrera 2008; Ishibuchi et al. 2005a, b; Karr and Gentry 1993), evolutionary algorithms have been successfully used for accuracy improvement of FRBSs in control, regression, and classification tasks. The main advantage of FRBSs is their linguistic interpretability; on the other hand, improvement in system accuracy is generally achieved at the expense of interpretability (Casillas et al. 2003b). While in the past a lot of attention has been paid only to system accuracy, in the last years an increasing number of papers have focused on a good trade-off between interpretability and accuracy (Casillas et al. 2005, 2007; Nauck and Kruse 1999; Yen et al.1998): in this framework, multi-objective evolutionary algorithms (MOEAs) (Coello Coello and Lamont 2004; Coello Coello 2006; Deb 2001; Zitzler et al. 2000) have proved to be very effective in searching for optimal solutions to problems that incorporate multiple performance criteria in competition with each other. MOEAs generate a family of equally valid solutions, where each solution tends to satisfy a criterion to a higher extent than another. Different solutions

P. Ducange · B. Lazzerini · F. Marcelloni (✉)
Dipartimento di Ingegneria dell'Informazione, Elettronica,
Informatica, Telecomunicazioni, University of Pisa,
Via Diotisalvi 2, 56122 Pisa, Italy
e-mail: f.marcelloni@iet.unipi.it

P. Ducange
e-mail: p.ducange@iet.unipi.it

B. Lazzerini
e-mail: b.lazzerini@iet.unipi.it

are compared with each other by using the notion of Pareto dominance. A solution **x** associated with a performance vector **u** dominates a solution **y** associated with a performance vector **v** if and only if, $\forall i \in \{1,\ldots,I\}$, with I the number of the criteria, u_i performs better than, or equal to, v_i and $\exists i \in \{1,\ldots,I\}$ such that u_i performs better than v_i, where u_i and v_i are the ith elements of vectors **u** and **v**, respectively. A solution is said to be Pareto-optimal if it is not dominated by any other possible solution. The set of Pareto-optimal solutions is denoted as Pareto front. Thus, the aim of a multi-objective search algorithm is to discover a family of solutions that are a good approximation of the Pareto front. In the case of multi-objective genetic fuzzy modeling, each solution in the front represents an FRBS with an associated trade-off between accuracy and interpretability.

In Ishibuchi et al. (1997) and Ishibuchi and Yamamoto (2004), MOEAs have been used for fuzzy rule selection, after extracting a large number of candidate rules from numerical data by a heuristic approach. The objective of these studies is to find non-dominated subsets of candidate fuzzy rules with respect to both accuracy maximization and complexity minimization. Furthermore, in Ishibuchi and Nojima (2007) and Cococcioni et al. (2007), the authors show how sets of fuzzy rule bases can be generated with different trade-offs between accuracy and complexity/interpretability. Each rule is coded as an integer string, with each integer representing a fuzzy set, and a rule set is represented in a chromosome by concatenating as many strings as there are rules. Ad-hoc genetic operators are implemented to generate the offspring populations.

In Cordon et al. (2003), a multi-objective genetic algorithm has been used to jointly perform feature selection and fuzzy set granularity learning in order to obtain FRBSs composed of a compact set of comprehensive fuzzy rules with high classification ability. Finally, in Alcalá et al. (2007), a new method is presented that, by considering selection of rules together with tuning of membership functions, achieves solutions only in the Pareto zone with the highest accuracy. These solutions have the lowest possible number of rules but still present the highest accuracy.

In the last years, FRBSs have been extensively applied in the framework of pattern recognition for their good characteristics of generalization and their capability to explain how they perform classification (Casillas et al. 2001; Chang and Lilly 2004; Chi et al. 1996; Cordon et al. 1999; Ho et al. 2004; Ishibuchi and Yamamoto 2005; Nakashima et al. 2007). These FRBSs are denoted as fuzzy rule-based classifiers (FRBCs) in the literature.

In this paper, we focus on binary classifiers, that is, classifiers that assign patterns to one of two different classes, typically denoted as positive and negative classes. We use FRBCs which exploit the maximum matching method as reasoning method (see Cordon et al. 1999) and the penalized certainty factor as weight (see Ishibuchi and Yamamoto 2005). We exploit a multi-objective evolutionary approach to generate FRBCs with rule bases characterized by different trade-offs between accuracy and interpretability.

Usually, the accuracy of a classifier is evaluated in terms of percentage of correct classifications, and the objective of a classifier identification process is to maximize this percentage (or, equivalently, to minimize the misclassification rate). This objective might not be appropriate for the application domains characterized by highly imbalanced distributions of patterns, with positive cases composing just a small fraction of the available data used to train the classifier, or when the cost of misclassification of the positive patterns is different from the cost of misclassification of the negative patterns.

As regards classification of imbalanced datasets, several techniques have been proposed to overcome the bias toward the majority class that can result from the classical machine learning algorithms (Batista et al. 2004; Chawla et al. 2002). In Fernandez et al. (2008), authors have studied the behavior of FRBCs for imbalanced datasets, focusing on the synergy between the preprocessing mechanism of instances and the configuration of the systems. Good results have been achieved by using the algorithm proposed in Chi et al. (1996) to identify the rule base after the preprocessing step. However, the metric used does not take the different misclassification costs of each class into account and the obtained FRBCs lose their typical interpretability due to the high number of rules.

In Nakashima et al. (2007), a cost-sensitive approach for image-processing tasks based on FRBCs has been presented: the FRBC has been modified to incorporate the concept of pattern weight, which can be considered as the cost of an input pattern being misclassified. Given an initial rule base, an iterative learning algorithm has been proposed to adjust the rule weights in order to minimize the FRBC misclassification cost.

The receiver operating characteristic (ROC) curve analysis has proved to be very effective to compare different binary classifiers (Fawcett 2003, 2006; Provost and Fawcett 2001). The ROC curve is usually obtained by varying the parameters representing the binary classifier, and, for each classifier so generated, by plotting a point in the plane of the sensitivity and of the complement to one of the specificity. The sensitivity and the specificity express, respectively, how well the system classifies patterns belonging to the positive class and to the negative class. Sensitivity and specificity should be concurrently optimized. In some previous works (Anastasio et al. 1998; Everson and Fieldsend 2006; Kupinski and Anastasio

1999) this optimization is performed by an evolutionary multi-objective approach for binary non-fuzzy classifiers. In this paper, we extend this approach to FRBCs and add a third objective, namely the system complexity. We measure the complexity as sum of the conditions which compose the antecedents of the rules included in the FRBC. Using a well-known multi-objective evolutionary algorithm, namely the NSGA-II (Deb et al. 2002), we manage to provide a set of binary classifiers with a good trade-off among complexity, sensitivity, and specificity. After the optimization process, all the classifiers obtained are projected onto the ROC plane: in order to select a set of potentially optimal classifiers we use the ROC convex hull method proposed in (Provost and Fawcett 2001).

We applied the proposed approach to 11 datasets from the UCI repository (Asuncion and Newman 2007) and to 2 real medical datasets. All the datasets are highly imbalanced. To evaluate the global performance of the set of classifiers, we used the area under the ROC convex hull (AUCH). Further, we performed the Wilcoxon signed-rank test using the AUCH as a metric to compare our approach with two-two-objective evolutionary approaches and one heuristic approach to FRBC generation, and with three well-known classifiers. Finally, we show the application of the proposed approach to the generation of concrete FRBCs for discriminating the regions of interest (ROIs) in the lungs as nodules and non-nodules in a Computer Aided Diagnosis (CAD) system, and for discriminating calcification (cancerous samples) and non calcification (non-cancerous samples) in mammography.

2 Fuzzy rule-based classifiers

Pattern classification consists in assigning a class C_j from a predefined class set $C = \{C_1,\ldots, C_K\}$ to an object, represented as an F-dimensional point in a feature space $\Re^F$. Let $\mathbf{X} = \{X_1,\ldots, X_F\}$ be the set of input variables and U_f, $f = 1,\ldots, F$, be the universe of the fth variable. Let $P_f = \{A_{f,1},\ldots,A_{f,T_f}\}$, $f = 1,\ldots,F$, be a fuzzy partition with T_f fuzzy sets of universe U_f. The mth rule R_m $(m = 1,\ldots,M)$ of an FRBC is typically expressed as:

$$R_m : \textbf{IF } X_1 \textbf{ is } A_{1,j_{m,1}} \textbf{ and} \ldots \textbf{and } X_F \textbf{ is } A_{F,j_{m,F}} \quad \textbf{THEN } Y \textbf{ is } C_{j_m} \text{with } RW^m \tag{1}$$

where Y is the classifier output, $C_{j_m} \in C$ is the class label associated with the mth rule, $j_{m,f} \in [1, T_f]$ identifies the index of the fuzzy set (among the T_f fuzzy sets of the partition P_f), which has been selected for X_f in the rule R_m, and RW ^{m} is the *rule weight*, i.e., a *certainty degree* of the classification in the class C_{j_m} for a pattern belonging to the fuzzy subspace delimited by the antecedent of the rule.

Let $T = \{(\mathbf{x}_1, y_1),\ldots, (\mathbf{x}_N, y_N)\}$ be a training set composed of N input–output $(\mathbf{x}_t, y_t)$ pairs, with $\mathbf{x}_t = [x_{t,1},\ldots,x_{t,F}] \in \Re^F$ and $y_t \in C$. The strength of activation (*matching degree* of the rule with the input) of the rule R_m is calculated as $w_m(\mathbf{x}_t) = \prod_{f=1}^{F} A_{f,j_{m,f}}(x_{t,f})$, and the *association degree* with the class C_{j_m} is calculated as $h_m(\mathbf{x}_t) = w_m(\mathbf{x}_t) \cdot \mathrm{RW}^m$.

In the last few years, different approaches have been proposed to calculate the value of the rule weight RW^m (see Cordon et al. 1999, Ishibuchi and Yamamoto 2005 and Mansoori et al. 2007).

Two of the most known approaches are

(1) *certainty factor* $CF_m = \frac{\sum_{\mathbf{x}_t \in C_{j_m}} w_m(\mathbf{x}_t)}{\sum_{t=1}^{N} w_m(\mathbf{x}_t)}$ (Cordon et al. 1999), where N is the total number of patterns;

(2) *penalized certainty factor* $P\_CF_m = CF_m - \frac{\sum_{x_t \notin C_{j_m}} w_m(\mathbf{x}_t)}{\sum_{t=1}^{N} w_m(\mathbf{x}_t)}$ (Ishibuchi and Yamamoto 2005).

A fuzzy rule-based classifier is also characterized by its *reasoning method*, which uses the information from the rule base to determine a class label for a specific input pattern.

In this paper, for the sake of simplicity, we focus on the maximum matching method as reasoning method and the penalized certainly factor as weight. In the maximum matching method, an input pattern is classified into the class corresponding to the rule with the maximum association degree calculated for the pattern: in case of tie, we randomly classify the pattern. The choice of this combination of reasoning method and rule weight is supported by the results shown in Fernandez et al. 2008, where an in-depth study on the use of FRBCs for imbalanced datasets has been performed. Here, different kinds of fuzzy rule reasoning methods, of rule weights and of fuzzy rule learning methods have been compared and a detailed analysis on the importance of the pre-processing in generating FRBCs using non-evolutionary approaches has been carried out. The authors have concluded that the FRBC which uses the maximum matching method as reasoning method and the penalized certainty factor as weight achieves the best results.

Once the reasoning method and the rule weight are selected, an FRBC can be completely described by the following matrix $J \in N^{M\times(F+1)}$

$$J = \begin{bmatrix} j_{1,1} & \cdots & j_{1,F} & C_{j_1} \\ \cdots & \cdots & \cdots & \cdots \\ j_{m,1} & \cdots & j_{m,F} & C_{j_m} \\ \cdots & \cdots & \cdots & \cdots \\ j_{M,1} & \cdots & j_{M,F} & C_{j_M} \end{bmatrix} \tag{2}$$

where the generic element (m, f) indicates that the fuzzy set $A_{f,j_{m,f}}$ has been selected for the variable X_f in the rule R_m, and C_{j_m} is the associated class label. As an example, let us consider a two input-single output system. Let $P_1 = \{A_{1,1}, A_{1,2}, A_{1,3}\}$, $P_2 = \{A_{2,1}, A_{2,2}, A_{2,3}\}$ be the uniform fuzzy partitions of the two input variables, and $C = \{C_1, C_2, C_3\}$ be the class set. Then, the following rule base

R_1: **IF** X_1 **is** $A_{1,2}$ **and** X_2 **is** $A_{2,1}$ **THEN** Y **is** C_1
R_2: **IF** X_1 **is** $A_{1,2}$ **and** X_2 **is** $A_{2,2}$ **THEN** Y **is** C_3
R_3: **IF** X_1 **is** $A_{1,1}$ **and** X_2 **is** $A_{2,1}$ **THEN** Y **is** C_2

is described by the matrix

$$J = \begin{bmatrix} 2 & 1 & 1 \\ 2 & 2 & 3 \\ 1 & 1 & 2 \end{bmatrix}. \qquad (3)$$

In this work, we deal with the binary classification problem and so $C = \{C_1, C_2\}$, where C_1 represents the negative class and C_2 the positive class.

After fixing the input variable partitions, the reasoning method, and the weight type, we identify a set of rule bases using an evolutionary approach in the multi-objective framework in order to obtain FRBCs that show simultaneously high accuracy and low complexity. Note that the approach described in the following sections can be exploited for FRBCs with any kind of weights, rules, and reasoning methods.

3 Three-objective evolutionary optimization in the ROC space

In the framework of binary classifiers, the ROC curve analysis (Fawcett 2006; Provost and Fawcett 2001) is typically adopted to compare different classifiers with each other. Let TPR = TP/(TP + FN) be the true positive rate and FPR = FP/(TN + FP) be the false positive rate, where TP = true positive (positive correctly classified), TN = true negative (negative correctly classified), FP = false positive (negative classified as positive), FN = false negative (positive classified as negative). TPR and FPR coincide, respectively, with the sensitivity and the complement to 1 of the specificity. A chart is built, where TPR is represented on the ordinate axis, and FPR on the abscissa axis, by varying the classifier parameters. In this way, a specific classifier is represented as a point corresponding to its pair (FPR, TPR). These pairs give origin to the so-called ROC curve. Typically, the curve is plotted by varying the classifier parameters by a fixed step. This approach is feasible when the number of parameters is very small; otherwise, the generation of possible combinations of parameter values would require too much computational effort.

To solve this problem, we started from the following consideration: one point in the ROC space is better than (dominates) another if it is located more north-west (higher TPR, lower FPR, or both). It follows that TPR and FPR are two conflicting objectives that have to be concurrently optimized within the identification process of a binary classifier. Moreover, in our case, we would like to determine comprehensible classifiers. As discussed in Ishibuchi and Yamamoto (2004), the comprehensibility of the FRBSs depends on four factors: (1) comprehensibility of the fuzzy partitions of the input variables, (2) simplicity of the fuzzy rules, (3) simplicity of the fuzzy rule base, and (4) simplicity of the fuzzy reasoning.

Comprehensibility of the fuzzy partitions basically relies on the linguistic interpretability of the fuzzy sets, which is related to their number and to the separation of the neighboring fuzzy sets. Simplicity of the fuzzy rules is related to the type of fuzzy rules and to the number of the inputs involved in each rule. Simplicity of the fuzzy rule base mainly depends on the number of the input variables and the number of rules. Simplicity of the fuzzy reasoning depends on the type of inference used to deduce conclusions from facts and rules.

In this paper, we do not take the first and fourth factors into account. As stated earlier, we assume that the universe of discourse of each variable is uniformly partitioned, and we adopt the maximum matching method and the general model as reasoning methods. As regards the second factor, we use rules which, adopting linguistic variables in the antecedent part and singleton values (the class label) in the consequent part, guarantee an optimal interpretability. As regards the third factor, in order to reduce the complexity of the FRBSs we restrict the number of rules. Actually, complexity can be computed as a sum of the conditions which compose the antecedents of the rules included in the classifiers.

Reducing complexity and improving specificity and sensitivity are often conflicting objectives. To deal with this antagonism and generate a set of potentially optimal FRBCs with different trade-offs among complexity, specificity, and sensitivity, we adopt an MOEA. MOEAs have been investigated by several authors in recent years (Coello Coello and Lamont 2004; Coello Coello 2006; Deb 2001; Zitzler et al. 2000). Some of the most popular MOEAs are the Strength Pareto Evolutionary Algorithm (SPEA) (Zitzler and Thiele 1999) and its evolution (SPEA2) (Zitzler et al. 2001), the Niched Pareto Genetic Algorithm (NPGA) (Horn et al. 1999) and the different versions of the Pareto Archived Evolution Strategy (Knowles and Corne 2002). Due to its performance and simplicity, one of the most popular MOEAs is certainly NSGA-II, which was proposed in Deb et al. (2002) as an improved version of the previous

Non-dominated Sorting Genetic Algorithm (Srinivas and Deb 1998). NSGA-II is a population-based MOEA, which uses an ad-hoc density-estimation metric and a non-dominance rank assignment. NSGA-II starts from an initial random population P_0 of N_{pop} individuals sorted based on the non-dominance. Each individual is associated with a rank equal to its non-dominance level (1 for the best level, 2 for the next-best level, and so on). More precisely, first the non-dominated individuals are found and associated with rank 1; then in order to find the individuals with rank 2, i.e., those in the next non-dominated front, the individuals of the first front are temporarily discarded, and so on. At each iteration t, $t = 0,\ldots, T_{\max}$, an offspring population Q_t of size N_{pop} is generated by selecting mating individuals through the binary tournament selection, and by applying crossover and mutation operators. Then the parent population P_t and the offspring population Q_t are combined so as to generate a new population0 $P_{\text{ext}} = P_t \cup Q_t$. A rank based on the non-dominance level is assigned to each individual in P_{ext}. Using these ranks, P_{ext} is split into different non-dominated fronts, one for each different rank. Within each front, a specific crowding measure, which represents the sum of the distances from each individual to its closest individual along each objective, is used to define an ordering among individuals. The new parent population P_{t+1} is generated by deleting from P_{ext} the worst N_{pop} individuals (considering first the ordering among the fronts and then among the individuals). The algorithm terminates when the number of iterations equals $T_{\max}$. In the following, we introduce the chromosome coding and the mating operators used in our approach.

3.1 Chromosome representation

To generate the chromosome representation, we first fix the number T_f, $f = 1,\ldots, F$, of the fuzzy sets which partition each linguistic variable X_f (note that the number of the fuzzy sets can be different from an input variable to another). Second, we create a uniform partition $P_f = \{A_{f,1},\ldots,A_{f,T_f}\}$ for each variable X_f, $f = 1,\ldots, F$. Let us assume that the fuzzy system consists of M rules. Then, the system is completely defined by the matrix $J \in N^{M\times(F+1)}$ shown in (2). Thus, each possible solution can be described through a chromosome composed of $M \cdot (F + 1)$ natural numbers, as in Fig. 1.

The chromosome in Fig. 2 represents the system described by matrix J in (3).

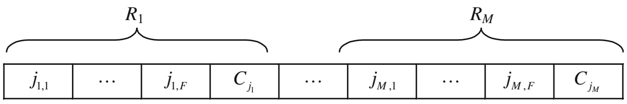

Fig. 1 Chromosome coding

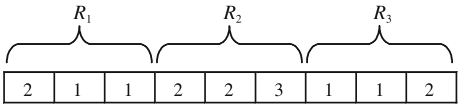

Fig. 2 An example of chromosome coding

As shown in (1), the antecedent of each rule involves all input variables. On the other hand some input variables could be irrelevant or even misleading in practical applications. Thus, it would be desirable to maintain only those input variables that actually contribute to characterize the system. To this aim, we add a new fuzzy set, denoted $A_{f,0}$, $f = 1,\ldots, F$, for each of the F input partitions P_f:$A_{f,0}$ is characterized by a membership function equal to 1 on the overall universe. This means that the condition X_f is $A_{f,0}$, denoted as *don't care condition* in the literature (Ishibuchi et al. 1997), does not affect the computation of the activation degree. In other words, for the specific rule, the variable X_f is not taken into account. The terms $A_{f,0}$ allow generating rules which contain only a subset of the input variables, thus implicitly performing a kind of feature selection. It follows that $j_{m,f} \in [0, T_f]$, $f = 1,\ldots,F$. A two-input fuzzy model with three rules can be described, for example, by the following matrix J:

$$J = \begin{bmatrix} 0 & 1 & 2 \\ 3 & 2 & 1 \\ 3 & 0 & 1 \end{bmatrix} \tag{4}$$

which corresponds to the rule base:

R_1: **IF and** X_2 **is** $A_{2,1}$ **THEN** Y **is** C_2
R_2: **IF** X_1 **is** $A_{1,3}$ **and** X_2 **is** $A_{2,2}$ **THEN** Y **is** C_1
R_3: **IF** X_1 **is** $A_{1,3}$ **THEN** Y **is** C_1

and to the chromosome in Fig. 3.

An input variable that is not included in at least one rule is eliminated. In this way we perform a selection of the inputs. In real applications, we are not interested in generating an approximation of the overall Pareto optimal front: too complex solutions with a large number of rules are not typically significant. Thus, to speed up the accuracy computation, we limit the search space by imposing a fixed maximum number $M_{\max}$ of rules which can compose the system.

Every chromosome is associated with an objective vector, each element of which typically expresses the fulfillment degree of a different objective. In our case, we use a 3D vector. The first element of the vector measures the

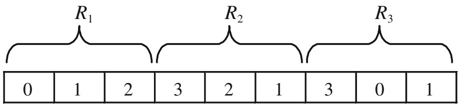

Fig. 3 Chromosome corresponding to matrix J in (4)

complexity as the number of genes corresponding to the antecedents which differ from 0, that is, the sum of the input variables actually used in each of the M rules. In the previous example, the complexity of the solution is 4 since we have one, two and one input variables for rules R_1, R_2 and R_3, respectively. The second and third elements express, respectively, TPR and FPR.

3.2 Genetic operators

Our implementation of NSGA-II uses the one-point crossover operator and three kinds of mutation operators.

Let c_1 and c_2 be the two solutions selected to generate two new individuals. The one-point crossover operator cuts the chromosomes c_1 and c_2 at some chosen common gene and swaps the resulting sub-chromosomes. The common gene is chosen by randomly extracting a number in $[M_{\min}, \rho_{\min}]$, where $M_{\min}$ is the minimum number of rules that must be present in a rule base, and $\rho_{\min}$ is the minimum number of rules in c_1 and c_2. We set $M_{\min} = 1$ in the experiments.

The first mutation operator adds γ rules to the rule base, where γ is randomly chosen in $\{1,\ldots, \gamma_{\max}\}$. The upper bound $\gamma_{\max}$ is fixed by the user. If $\gamma + M > M_{\max}$, then $\gamma = M_{\max} - M$. For each rule m added to the chromosome, we generate n natural random numbers between 1 and F to determine the input variables which compose the antecedent part of the rule. Finally, for each selected input variable f, we generate a random natural number $j_{m,f}$ between 1 and T_f, which determines the fuzzy set $A_{f,j_{m,f}}$ to be used in the antecedent of the rule m. We guarantee at least one non-zero condition is present in the antecedent part. The class label of the rule is assigned according to the maximum certainty factor of the training patterns of each class.

The second mutation operator removes λ rules from the rule base, where λ is selected at random in $\{1,\ldots, \lambda_{\max}\}$. In the experiments, we used $\lambda_{\max} = \min(\phi_{\max}, M - M_{\min})$, where $\phi_{\max}$ is fixed by the user and M is the number of rules of the individual.

The third mutation operator randomly changes δ antecedent conditions of matrix J. The number δ is randomly generated in $\{1,\ldots, \delta_{\max}\}$. The upper bound $\delta_{\max}$ depends on the complexity of the selected solution to be muted. We fixed $\delta_{\max}$ to the 10% of the complexity. We ensure that, when the operator is selected, at least one condition is modified, even if the complexity is lower than 10. The class label of the rule that has been modified is assigned according to the maximum certainty factor of the training patterns of each class. Each time a rule is added or modified, the corresponding weight is updated.

The probability of applying the crossover operator is 0.5. When the application of the crossover operator is selected, the mutation is applied with probability 0.01; otherwise it is always applied. When the application of the mutation is selected, the first and the second mutation operators are applied with the same probability of 0.4; otherwise the third mutation operator is applied.

We experimentally verified that the crossover operator helps the algorithm perform a good exploitation of the solutions generated so far. Furthermore, the first mutation operator is useful to generate solutions with high complexity, the second operator is useful to reduce the complexity of the solutions, and the third operator allows performing a good exploitation of the search space.

4 The ROC convex hull technique

The output of the NSGA-II execution is a three-dimensional Pareto front. First of all, we project the points of the Pareto front onto the TPR–FPR plane, so as to obtain a ROC curve. The ROC curve shows the behavior of a classifier without any reference to the class distribution or the classification error cost. This means that the ROC analysis can be appropriate to choose the best global classifier only if there exists a classifier that dominates all the others over the overall ROC space. Otherwise, the answer to the question about which classifier is the best depends on both the class distribution and the classification error cost. The expected cost of applying a classifier (Provost and Fawcett 2001) represented by a point (FP,TP) in the ROC space can be expressed as follows:

$$\text{cost} = \text{FPR} \cdot P(n) \cdot c(Y, n) + \text{FNR} \cdot P(p) \cdot c(N, p)$$

where $P(p)$ and $P(n) = 1 - P(p)$ are the a priori probabilities of a positive example and a negative example, respectively, $c(Y, \text{n})$ and $c(N, \text{p})$ are, respectively, the false-positive cost and the false-negative cost, and $\text{FNR} = 1 - \text{TPR} = 1 - \text{TP}/(\text{TP} + \text{FN}) = \text{FN}/(\text{TP} + \text{FN})$ is the false-negative rate. In order to solve the problem of choosing the best global classifier, we can resort to the ROC convex hull technique (Provost and Fawcett 2001). We recall that a classifier is potentially optimal if and only if it lies on the convex hull of the set of points in the ROC space. The convex hull technique allows choosing locally optimum classifiers, that is, the classifiers that are optimal with respect to given class distributions and classification error costs.

Recalling the expression of the expected classification cost by the classifier represented by the point (FPR, TPR) in the ROC space, once fixed the values for $P(n)$, $c(Y, n)$, $P(p)$ and $c(N, \text{p})$, we obtain a family of parallel lines (called iso-cost lines) with slope$\frac{P(n) \cdot c(Y,n)}{P(p) \cdot c(N,p)}$.

The points belonging to the same line have the same cost, and the cost decreases as we move to parallel lines closer to the point (0,1), i.e., more north-west. For a given ROC curve, the point that minimizes the classification cost is the tangent point between the ROC convex hull and the family of parallel lines (local optimum condition) as shown in Fig. 4. This approach is robust because it scales gracefully to any degree of precision in specifying the classification error costs and the class distributions without requiring to perform a new FRBC identification process.

To evaluate the goodness of the convex hull and, therefore, the global performance of the set of classifiers in the convex hull, we compute the AUCH (Fawcett 2006). Since the AUCH is a portion of the unit square, its value will always be between 0 and 1. Anyway, because random classifiers lie on the diagonal line between (0,0) and (1,1), with an area of 0.5, no realistic classifier should have an AUCH smaller than 0.5. The AUCH results to be very effective to compare different types of classifiers. Indeed, we can generate the ROC convex hull for each type of classifier to be compared (by adopting the approach described in this paper or varying appropriately the parameters which define the type of classifier) and then compute the AUCH for each convex hull. The type of classifier with the highest AUCH will be the one which on average (with different values of the classification error costs and class distributions) outperforms the others. We will adopt the AUCH in order to assess different types of classifiers in the experimental part.

5 Experimental results

We tested our method on 11 datasets from the UCI repository (Asuncion and Newman 2007) and 2 real medical datasets, namely Nodules (we will discuss this dataset in the following) and Mammography datasets (Woods et al. 1993). These datasets are imbalanced datasets because the classification categories are not equally represented. Real-world datasets often consist of a large percentage of *normal* (negative class) examples and a small percentage of *abnormal* or *interesting* examples (positive class). Table 1 summarizes the characteristics of the 13 datasets.

Various techniques have been proposed to deal with imbalanced datasets and two main types of approaches are adopted:

(1) a data level-approach, which consists of balancing the class distribution by either over-sampling the minority class, or under-sampling the majority class, or using a hybrid technique (Batista et al. 2004; Chawla et al. 2002);

(2) an algorithmic level-approach, which, for example, adjusts the costs per class (Provost and Fawcett 2001) or learns from just one class (recognition based learning) instead of learning from the two classes (discrimination based learning) (Raskutti and Kowalczyk 2004).

In this work, we follow the algorithmic level-approach: we maximize TPR and minimize FPR at the same time, and then we select the best classifier using the convex hull method.

In the following, we first (Sect. 5.1) compare our three-objective evolutionary approach with two-two-objective evolutionary approaches aimed at identifying a set of FRBCs with different trade-offs between accuracy and complexity. We applied the two-objective approaches both on the original and rebalanced datasets. Then, in Sect. 5.2, we compare our approach with three well-known classifiers, namely the linear discriminant classifier (ldc), the naïve bayes classifier (naïve) and the C4.5 rule learning (C4.5), and a heuristic method proposed in Chi et al. (1996) to generate FRBCs. We denote this method as Chi FRBC in

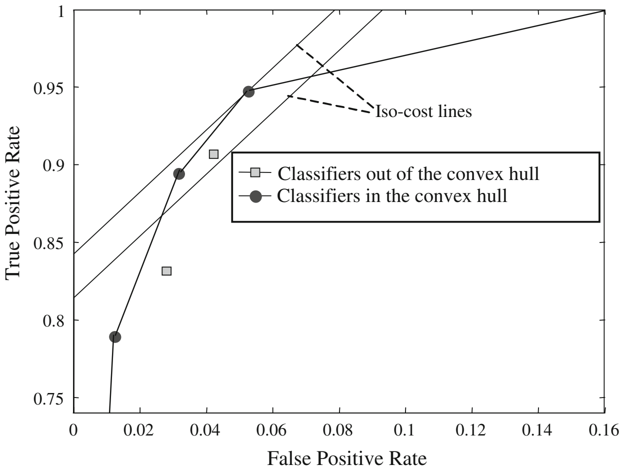

Fig. 4 Selection of the best classifier in the ROC space

Table 1 Dataset characteristics

Dataset	Examples	Features	% Positive	% Negative
ecoli-0-1-3-7_vs_2-6	281	7	2.49	97.51
shuttle0vs4	1,829	9	6.72	93.28
yeastB1vs7	459	8	6.72	93.28
shuttle2vs4	129	9	4.65	95.35
glass-0-1-6_vs_2	192	9	8.89	91.11
glass-0-1-6_vs_5	184	9	4.89	95.11
page-blocks-1-3_vs_4	472	10	5.93	94.07
yeast-0-5-6-7-9_vs_4	528	8	9.66	90.34
yeast-1-2-8-9_vs_7	947	8	3.17	96.83
yeast-1-4-5-8_vs_7	693	8	4.33	95.67
yeast-2_vs_4	514	8	9.92	90.08
nodules	1,984	4	3.20	96.80
mammography	7,893	6	3.01	96.99

Table 2 The types of classifiers used in the experiments

Type of Classifier	Acronym
FRBC generated by the three-objective evolutionary approach	3ob_FRBC
FRBC generated by the two-objective evolutionary approach (with AUC as accuracy)	2ob_ FRBC_AUC
FRBC generated by the two-objective evolutionary approach (with GM as accuracy)	2ob_FRBC_GM
FRBC generated by the two-objective evolutionary approach (with AUC as accuracy and Smote as data rebalancing technique)	2ob_FRBC_AUC_Sm
FRBC generated by the two-objective evolutionary approach (with GM as accuracy and Smote as data rebalancing technique)	2ob_FRBC_GM_Sm
FRBC generated by Chi et al. algorithm	Chi FRBC
Linear Discriminant Classifier	ldc
Naïve Bayes	naïve
C4.5 Learning Rule	C4.5

the following. Table 2 shows the different types of classifiers we employed in the experiments and the acronyms used to refer to them. For both the comparisons, we performed the Wilcoxon signed-rank test (Sheskin 2003 and Wilcoxon 1945) for pair-wise statistical assessment of the different algorithms. This test is based on computing the differences on two sample means (typically, accuracy obtained by a pair of different algorithms on different datasets). In our case, we adopt the AUCH as metric for the statistical test, so as to provide a global evaluation of the different approaches. To perform the test, we use a level of confidence $\alpha = 0.05$.

Finally, in Sect. 5.3 once fixed the misclassification costs and the class distribution probabilities, we discuss two examples of selection of the most suitable classifiers for, respectively, the Nodules and Mammography datasets. We performed the selection step only for these datasets because only for these datasets we can approximately fix meaningful misclassification costs.

In all the experiments, we adopted a fivefold cross-validation model, i.e., we split randomly each data set into five folds, each containing the 20% of the patterns and used four folds for training and one for testing.

5.1 Comparison among different evolutionary approaches

To show the advantages of using TPR and FPR as objectives of an evolutionary multi-objective optimization for the design of binary FRBCs, we executed the NSGA-II algorithm with the same parameters, operators, and probabilities as in the proposed approach, but considering a unique objective as accuracy measure. We used both the area under the ROC curve (AUC) and the geometric mean (GM). Both GM and AUC combine together TPR and FPR, in terms of product and sum, respectively. Indeed $\mathrm{GM} = \sqrt{\mathrm{TPR} \cdot \mathrm{TNR}} = \sqrt{\mathrm{TPR} \cdot (1 - \mathrm{FPR})}$ and $\mathrm{AUC} = (1 + \mathrm{TPR} - \mathrm{FPR})/2$.

We fixed the population size of NSGA-II to 150, and $\gamma_{\max}$ and $\phi_{\max}$ to 5. We adopted this value for the last two parameters supported by the following consideration. Values larger than 5 might generate an offspring too different from the parents, thus making mutation operators too much oriented to exploration rather than exploitation. We verified experimentally that in this case the convergence of the algorithm is slower. Fixing the values of the two parameters to 5 guarantees a good trade-off between exploration and exploitation. Indeed, if the algorithm chooses randomly a value between 1 and 2, we have a good exploitation, whereas if the algorithm chooses randomly a value between 4 and 5, we have a good exploration. For each fold, we executed NSGA-II for 300 iterations. To guarantee a good level of interpretability, we fixed the maximum number $M_{\max}$ of rules to 20 and we uniformly partitioned the universe of each variable using five fuzzy sets labeled LOW, MEDIUM LOW, MEDIUM, MEDIUM HIGH and HIGH.

As regards the two-objective evolutionary approaches, we performed simulations using both the original (imbalanced) and the rebalanced datasets (we employed the SMOTE over-sampling algorithm proposed in Chawla et al. (2002) to rebalance the datasets). For each solution in the final approximated Pareto fronts, we computed the TPR and the FPR, and plotted the corresponding (TPR, FPR) point in the ROC space. Finally, we extracted the convex hull in the ROC space both on training and test sets. To evaluate the goodness of the set of the solutions in the convex hull we calculated the AUCH.

Tables 3, 4 and 5 show the average results obtained on all the datasets by our approach, and by the two-objective evolutionary approaches without and with rebalancing, respectively. In the tables, $AUCH_{tr}$ and $AUCH_{ts}$ represent the average values of the AUCH on training and test sets, respectively, C_{tr} and C_{ts} represent the average values of complexity of the most complex solutions of the ROC convex hull on the training and test sets, respectively, and C_{max} represents the maximum possible value of complexity for each dataset (calculated as the product between M_{max} and the number of features of the dataset). In the last line of each table, we also report in bold the average results obtained on all the datasets.

Tables 3 and 4 show that, on average, the three-objective evolutionary approach outperforms the two-objective evolutionary approaches both on the training and test sets with respect to the AUCH metric. Table 5 highlights the fact that the rebalancing process improves the performance of the two-objective approaches, but cannot contrast the three-objective approach which seems to explore the search space more effectively. We can observe that all the approaches generate FRBCs characterized by a value of complexity much lower than the possible maximum value C_{max}. This confirms the effectiveness of the multi-objective approach in concurrently maximizing accuracy and minimizing complexity. To assess whether there exist significant differences between the results achieved by our three-objective approach and each of the other two-objective approaches, we adopted the Wilcoxon signed-rank test. Table 6 shows the results of the test application. Here, R^+ and R^- represent the sums of the ranks corresponding to our approach and to the other evolutionary approaches, respectively. The null hypothesis is accepted if p value $> \alpha$. We note that the null hypothesis is always rejected in favor of the three-objective approach, thus confirming that the three-objective evolutionary approach always outperforms, in terms of AUCH on the test set, the two-objective evolutionary approaches.

5.2 Comparisons with non-evolutionary classifiers

In this section, we compare our three-objective evolutionary approach with three well-known classifiers (ldc, naïve and C4.5) and the Chi FRBC. We used the versions of ldc and naïve implemented in the PRTools package for MATLAB (Duin 2007) and the version of C4.5 implemented in the software Keel (Alcalá-Fdez et al. 2009).

As regards ldc and naïve, we first generated the ROC curves, both on training and test sets, by varying the output threshold from zero to one, with step 0.005, and then we extracted the points of the convex hull. For the Chi FRBC and C4.5, we have only one point in the ROC curve (in particular, for C4.5 this point was obtained by using the default parameters of the Keel software). Thus, as discussed in Fawcett (2006), we considered the convex hull as composed by this point, and the points (0,0) and (1,1).

Note that, both for the ldc, naïve, and C4.5 classifiers, and for the Chi FRBC, we rebalanced only the training sets by using SMOTE and evaluated the generalization capability on never-shown patterns using the imbalanced test sets (which are exactly the same as used in the evolutionary generated FRBCs).

Table 7 shows the average results obtained on all the datasets by the Chi FRBC, ldc, naïve and C4.5.

We note that the best average results among the non-evolutionary classifiers are achieved by C4.5 on the training set and by ldc on the test set. In any case, our

Table 3 Results achieved by the three-objective evolutionary approach

	3ob_FRBC				C_{max}
Dataset	$AUCH_{tr}$	$AUCH_{ts}$	C_{tr}	C_{ts}	
ecoli-0-1-3-7_vs_2-6	1.00 ± 0.00	0.89 ± 0.22	27.00 ± 7.65	10.20 ± 4.49	140
shuttle0vs4	1.00 ± 0.00	1.00 ± 0.00	11.40 ± 6.23	10.40 ± 4.16	180
yeastB1vs7	0.95 ± 0.01	0.90 ± 0.04	82.40 ± 14.05	25.60 ± 12.30	160
shuttle2vs4	1.00 ± 0.00	1.00 ± 0.00	10.20 ± 0.84	10.00 ± 1.22	180
glass-0-1-6_vs_2	0.96 ± 0.02	0.91 ± 0.08	52.80 ± 5.17	25.40 ± 16.15	180
glass-0-1-6_vs_5	1.00 ± 0.00	0.98 ± 0.02	21.00 ± 11.42	13.60 ± 10.11	180
page-blocks-1-3_vs_4	1.00 ± 0.00	1.00 ± 0.00	35.40 ± 11.24	21.40 ± 16.56	200
yeast-0-5-6-7-9_vs_4	0.96 ± 0.01	0.93 ± 0.05	96.20 ± 19.94	50.40 ± 20.55	160
yeast-1-2-8-9_vs_7	0.90 ± 0.02	0.85 ± 0.05	70.40 ± 14.40	47.60 ± 23.80	160
yeast-1-4-5-8_vs_7	0.86 ± 0.02	0.83 ± 0.04	114.40 ± 18.77	55.80 ± 36.98	160
yeast-2_vs_4	0.99 ± 0.00	0.99 ± 0.01	74.60 ± 27.41	18.40 ± 8.20	160
nodules	0.97 ± 0.00	0.97 ± 0.02	51.40 ± 11.22	22.20 ± 16.05	80
mammography	0.96 ± 0.01	0.95 ± 0.03	104.00 ± 3.08	97.60 ± 9.21	120
	0.97 ± 0.01	**0.94 ± 0.04**	**57.78 ± 10.88**	**31.43 ± 13.83**	**158.46**

Table 4 Results achieved by the two-objective evolutionary approaches without rebalancing

Dataset	2ob_ FRBC_AUC				2ob_FRBC_GM				C_{max}
	$AUCH_{tr}$	$AUCH_{ts}$	C_{tr}	C_{ts}	$AUCH_{tr}$	$AUCH_{ts}$	C_{tr}	C_{ts}	
ecoli-0-1-3-7_vs_2-6	0.98 ± 0.02	0.84 ± 0.21	20.40 ± 8.41	19.60 ± 9.42	0.98 ± 0.02	0.89 ± 0.19	19.20 ± 5.93	15.40 ± 4.88	140
shuttle0vs4	1.00 ± 0.00	1.00 ± 0.00	11.20 ± 1.30	11.20 ± 1.30	1.00 ± 0.00	1.00 ± 0.00	12.00 ± 1.87	12.00 ± 1.87	180
yeastB1vs7	0.86 ± 0.02	0.75 ± 0.04	65.20 ± 29.59	23.00 ± 15.13	0.85 ± 0.03	0.70 ± 0.07	60.00 ± 25.27	40.20 ± 18.91	160
shuttle2vs4	1.00 ± 0.00	1.00 ± 0.00	11.20 ± 1.30	11.20 ± 1.30	1.00 ± 0.00	0.99 ± 0.01	11.20 ± 1.48	11.20 ± 1.48	180
glass-0-1-6_vs_2	0.83 ± 0.05	0.70 ± 0.10	38.80 ± 26.80	32.20 ± 15.96	0.74 ± 0.15	0.69 ± 0.16	49.20 ± 48.95	29.00 ± 23.53	180
glass-0-1-6_vs_5	0.90 ± 0.22	0.87 ± 0.21	20.60 ± 11.13	17.80 ± 12.09	0.90 ± 0.22	0.84 ± 0.22	19.80 ± 16.95	17.40 ± 18.11	180
page-blocks-13_vs_4	0.98 ± 0.03	0.98 ± 0.03	18.80 ± 4.76	17.20 ± 6.83	1.00 ± 0.00	0.99 ± 0.01	30.40 ± 14.74	28.20 ± 15.93	200
yeast-0-5-6-7-9_vs_4	0.87 ± 0.03	0.82 ± 0.07	80.20 ± 12.11	57.60 ± 21.76	0.87 ± 0.03	0.82 ± 0.05	75.20 ± 27.52	46.60 ± 22.53	160
yeast-1-2-8-9_vs_7	0.72 ± 0.04	0.70 ± 0.08	31.40 ± 25.01	27.60 ± 19.93	0.67 ± 0.11	0.65 ± 0.14	77.20 ± 50.51	77.20 ± 50.51	160
yeast-1-4-5-8_vs_7	0.68 ± 0.10	0.61 ± 0.07	51.80 ± 37.29	29.60 ± 25.70	0.65 ± 0.09	0.56 ± 0.07	47.00 ± 48.69	15.00 ± 12.96	160
yeast-2_vs_4	0.90 ± 0.05	0.86 ± 0.06	54.60 ± 25.12	53.00 ± 27.82	0.92 ± 0.02	0.87 ± 0.04	35.80 ± 24.85	23.80 ± 12.21	160
nodules	0.90 ± 0.01	0.83 ± 0.06	27.00 ± 9.03	20.00 ± 11.20	0.90 ± 0.02	0.86 ± 0.09	15.60 ± 6.62	14.20 ± 6.38	80
mammography	0.90 ± 0.02	0.88 ± 0.03	67.00 ± 14.27	62.60 ± 17.95	0.90 ± 0.00	0.89 ± 0.04	65.40 ± 18.23	52.80 ± 15.27	120
	0.89 ± 0.05	**0.83 ± 0.07**	**38.32 ± 16.81**	**29.43 ± 13.64**	**0.88 ± 0.05**	**0.83 ± 0.08**	**40.25 ± 22.43**	**28.14 ± 15.73**	**156.46**

Table 5 Results achieved by the two-objective evolutionary approaches with Smote rebalancing

Dataset	2ob_ FRBC_AUC_Sm				2ob_FRBC_GM_Sm				C_{max}
	$AUCH_{tr}$	$AUCH_{ts}$	C_{tr}	C_{ts}	$AUCH_{tr}$	$AUCH_{ts}$	C_{tr}	C_{ts}	
ecoli-0-1-3-7_vs_2-6	0.98 ± 0.01	0.83 ± 0.21	36.00 ± 14.71	4.60 ± 3.85	0.98 ± 0.01	0.86 ± 0.20	45.20 ± 15.22	25.20 ± 27.29	140
shuttle0vs4	1.00 ± 0.00	1.00 ± 0.00	5.00 ± 1.41	4.40 ± 1.52	1.00 ± 0.00	1.00 ± 0.00	4.80 ± 1.79	4.80 ± 1.79	180
yeastB1vs7	0.78 ± 0.07	0.69 ± 0.13	37.00 ± 13.78	16.00 ± 10.91	0.83 ± 0.02	0.77 ± 0.07	65.00 ± 15.62	40.20 ± 33.68	160
shuttle2vs4	1.00 ± 0.00	1.00 ± 0.01	23.00 ± 12.79	9.40 ± 3.36	1.00 ± 0.00	1.00 ± 0.00	11.80 ± 4.66	10.80 ± 2.49	180
glass-0-1-6_vs_2	0.81 ± 0.07	0.71 ± 0.14	57.40 ± 25.58	43.00 ± 21.64	0.80 ± 0.04	0.67 ± 0.12	78.20 ± 17.98	43.00 ± 25.82	180
glass-0-1-6_vs_5	0.95 ± 0.05	0.94 ± 0.03	45.20 ± 12.79	45.20 ± 12.79	0.95 ± 0.03	0.95 ± 0.04	39.40 ± 6.31	31.00 ± 10.22	180
page-blocks-1-3_vs_4	0.92 ± 0.08	0.92 ± 0.06	24.80 ± 15.74	21.80 ± 14.50	0.99 ± 0.01	0.98 ± 0.01	45.00 ± 27.46	28.20 ± 32.41	200
yeast-0-5-6-7-9_vs_4	0.84 ± 0.03	0.82 ± 0.05	63.40 ± 17.98	35.80 ± 25.05	0.84 ± 0.03	0.79 ± 0.08	68.20 ± 18.23	50.60 ± 13.20	160
yeast-1-2-8-9_vs_7	0.98 ± 0.01	0.83 ± 0.21	36.00 ± 14.71	4.60 ± 3.85	0.98 ± 0.01	0.86 ± 0.20	45.20 ± 15.22	25.20 ± 27.29	160
yeast-1-4-5-8_vs_7	0.75 ± 0.02	0.62 ± 0.08	87.00 ± 25.37	73.80 ± 28.35	0.75 ± 0.05	0.67 ± 0.03	94.40 ± 20.26	56.80 ± 30.80	160
yeast-2_vs_4	0.90 ± 0.03	0.89 ± 0.03	44.80 ± 21.79	34.00 ± 22.10	0.92 ± 0.02	0.89 ± 0.05	49.80 ± 21.81	19.40 ± 15.32	160
nodules	0.92 ± 0.01	0.88 ± 0.02	40.00 ± 7.65	23.20 ± 6.46	0.92 ± 0.01	0.89 ± 0.04	31.00 ± 15.12	25.40 ± 15.66	80
mammography	0.90 ± 0.04	0.86 ± 0.04	61.80 ± 35.52	60.80 ± 35.12	0.92 ± 0.02	0.88 ± 0.02	64.80 ± 14.41	53.20 ± 23.30	120
	0.90 ± 0.03	**0.85 ± 0.08**	**43.18 ± 16.91**	**28.97 ± 14.58**	**0.91 ± 0.02**	**0.86 ± 0.07**	**49.45 ± 14.93**	**31.83 ± 19.94**	**158.46**

three-objective approach outperforms all the approaches on the training set and on the test set, except for ldc which shows the same value of mean AUCH on the test set.

Table 8 shows the results of the application of the Wilcoxon signed-rank test between our three-objective evolutionary approach and the four non-evolutionary approaches. We observe that our approach statistically outperforms, in terms of AUCH on the test set, all the non-evolutionary classifiers except for ldc. Since in this case the null hypothesis is accepted, we conclude that there do not exist statistical differences between the results achieved by our approach and ldc. On the other hand, we have to consider that ldc is not interpretable.

5.3 Classifier selection for Nodules and Mammography datasets

In this section, once the misclassification costs and the class distribution probabilities are fixed, we present two examples of classifier selection for the Nodules and Mammography datasets, respectively. We performed the selection step just for these two real medical datasets, because only for these datasets we could approximately fix the misclassification costs. We extract the best classifier from the ROC convex hull generated by using the patterns of the training set and then assess the generalization capability of the selected classifier on the test set.

Table 6 Wilcoxon signed-rank test between the 3ob_FRBC and the four-two-objective evolutionary approaches

Comparison	R^+	R^-	Hypothesis ($\alpha = 0.05$)	*p* value
3ob_FRBC vs. 2ob_ FRBC_AUC	89.5	1.5	Rejected for 3ob_FRBC	0.003
3ob_FRBC vs. 2ob_FRBC_GM	89.5	1.5	Rejected for 3ob_FRBC	0.003
3ob_FRBC vs. 2ob_FRBC_AUC_Sm	89.5	1.5	Rejected for 3ob_FRBC	0.003
3ob_FRBC vs. 2ob_FRBC_GM_Sm	86.5	4.5	Rejected for 3ob_FRBC	0.004

Table 7 Results achieved by the four non-evolutionary approaches

Dataset	Chi FRBC		ldc		naïve		C4.5	
	$AUCH_{tr}$	$AUCH_{ts}$	$AUCH_{tr}$	$AUCH_{ts}$	$AUCH_{tr}$	$AUCH_{ts}$	$AUCH_{tr}$	$AUCH_{ts}$
ecoli-0-1-3-7_vs_2-6	0.98 ± 0.01	0.71 ± 0.01	0.98 ± 0.01	0.97 ± 0.06	0.89 ± 0.22	0.88 ± 0.22	0.98 ± 0.01	0.81 ± 0.22
shuttle0vs4	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
yeastB1vs7	0.80 ± 0.02	0.84 ± 0.00	0.88 ± 0.01	0.90 ± 0.03	0.90 ± 0.02	0.88 ± 0.04	0.91 ± 0.02	0.65 ± 0.05
shuttle2vs4	0.98 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	0.99 ± 0.01
glass-0-1-6_vs_2	0.72 ± 0.04	0.57 ± 0.09	0.87 ± 0.02	0.91 ± 0.06	0.88 ± 0.01	0.75 ± 0.10	0.95 ± 0.01	0.63 ± 0.11
glass-0-1-6_vs_5	0.97 ± 0.01	0.96 ± 0.00	0.96 ± 0.00	0.97 ± 0.03	0.99 ± 0.00	0.98 ± 0.03	0.97 ± 0.01	0.81 ± 0.24
page-blocks-1-3_vs_4	0.95 ± 0.01	0.96 ± 0.00	0.96 ± 0.00	0.96 ± 0.02	0.98 ± 0.00	0.97 ± 0.02	0.99 ± 0.00	0.99 ± 0.01
yeast-0-5-6-7-9_vs_4	0.82 ± 0.02	0.80 ± 0.01	0.88 ± 0.02	0.90 ± 0.05	0.90 ± 0.01	0.87 ± 0.04	0.91 ± 0.01	0.75 ± 0.09
yeast-1-2-8-9_vs_7	0.84 ± 0.02	0.86 ± 0.03	0.98 ± 0.01	0.97 ± 0.06	0.89 ± 0.22	0.88 ± 0.22	0.93 ± 0.03	0.63 ± 0.05
yeast-1-4-5-8_vs_7	0.75 ± 0.01	0.68 ± 0.11	0.77 ± 0.01	0.78 ± 0.04	0.85 ± 0.02	0.82 ± 0.07	0.94 ± 0.02	0.54 ± 0.03
yeast-2_vs_4	0.89 ± 0.02	0.86 ± 0.03	0.96 ± 0.00	0.97 ± 0.01	0.87 ± 0.21	0.88 ± 0.21	0.95 ± 0.01	0.88 ± 0.07
nodules	0.86 ± 0.02	0.82 ± 0.05	0.95 ± 0.01	0.95 ± 0.03	0.95 ± 0.00	0.94 ± 0.04	0.94 ± 0.01	0.87 ± 0.03
mammography	0.90 ± 0.00	0.86 ± 0.04	0.92 ± 0.01	0.90 ± 0.04	0.93 ± 0.01	0.91 ± 0.04	0.92 ± 0.02	0.86 ± 0.06
	0.88 ± 0.01	**0.84 ± 0.03**	**0.93 ± 0.01**	**0.94 ± 0.03**	**0.93 ± 0.06**	**0.90 ± 0.08**	**0.95 ± 0.01**	**0.80 ± 0.07**

5.3.1 Nodules dataset

Lung cancer represents the leading cause of death for malignancy in the world. Thus, automatic detection of malignant nodules is an important research topic. The advent of multi-slice, spiral computed tomography (CT), able to perform a 3D reconstruction of the interior of the human body, has opened new horizons for early detection of lung cancers (Awai et al. 2004).

We tested our approach using a dataset made of features extracted from 3D ROIs, which represent candidates for lung nodules. These ROIs are automatically extracted from lung multi-slice CT scans (all with medical report) (Antonelli et al. 2006), each consisting, on average, of 300 slices, each slice being a 512 × 512 pixel matrix. The slice thickness is 1.25 mm and the pixel size is 0.6 mm.

Table 8 Wilcoxon signed-rank test between the 3ob_FRBC and the non-evolutionary classifiers

Comparison	R^+	R^-	Hypothesis ($\alpha = 0.05$)	*p* value
3ob_FRBC vs. Chi FRBC	86.5	4.5	Rejected for 3ob_FRBC	0.005
3ob_FRBC vs. ldc	61.0	30.0	Accepted	0.514
3ob_FRBC vs. naïve	79.0	12.0	Rejected for 3ob_FRBC	0.021
3ob_FRBC vs. C4.5	90.5	0.5	Rejected for 3ob_FRBC	0.002

Each ROI may be either a nodule (positive class) or a non-nodule (negative class), e.g., a blood vessel. Specifically, we used a dataset $X = X_n \cup X_v$ made of 1,984 3D ROIs, divided into 64 nodules (set X_n), corresponding to $\cong$3.2% of the total, and 1,920 blood vessels (set X_v), corresponding to $\cong$96.8% of the total.

All the data used in the experiments are represented as patterns in $\Re^4$, after extracting the following features from the single 3D ROIs: sphericity, average gray level, standard deviation of the average gray level, and inverse of elongation. In particular, sphericity s is defined as $s = \frac{V(R_i \cap S_i)}{V(R_i)}$ where $V(\cdot)$, R_i, S_i are, respectively, the volume, the 3D ROI and the sphere equivalent to R_i [i.e., the sphere with center in the barycenter of R_i and radius equal to that of the sphere with volume $V(R_i)$]. The inverse of elongation $\frac{1}{el}$ is computed as $\frac{1}{\mathrm{el}} = \frac{8\left[\max_{j \in J(R_i)} (d_j)\right]^2}{\sum_{j \in J(R_i)} A(Z_j)}$, where $A(\cdot)$ is the area of the jth 2D ROI Z_j forming the 3D ROI R_i, $\max(\cdot)$ is the maximum, d_j is the thickness of Z_j (that is, d_j is the number of erosion steps necessary to delete Z_j), and $J(R_i)$ is the set of indexes j of the 2D ROIs composing the 3D ROI R_i.

Figure 5 shows the Pareto front approximations provided by NSGA-II for each fold. We projected the three-dimensional fronts onto the ROC plane and we represented the complexity with different gray levels (we associated higher values of complexity with darker gray levels). We also plotted a piecewise line that represents the ROC convex hull. As expected, a good trade-off between complexity and accuracy was obtained: we observe that classifiers characterized by high accuracy (high TPR and low FPR) are generally more complex than classifiers with low accuracy.

As an example of classifier selection, we fixed the a-priori probabilities $P(n)$ and $P(p)$ of the negative and positive classes to the class frequencies in the data set, that is, $P(n)$ equal to 1,920/1,984 ($\cong$97%) and $P(p)$ equal to 64/1,984 ($\cong$3%). Then, we chose a ratio $c(Y, \mathrm{n})/c(N, \mathrm{p})$ equal to 1/60. This ratio was motivated by considering the cost of a false negative much higher than the cost of a false positive. On the other hand, a false negative does not activate the adequate treatment and therefore might be extremely hazardous for the patient's health. We, therefore, obtained a slope for the optimal iso-cost lines equal to 0.5. Considering the cost of a false positive equal to one, the cost of a classifier can be calculated as follows: $\mathrm{cost} = \mathrm{FPR} \cdot 0.97 \cdot 1 + \mathrm{FNR} \cdot 0.03 \cdot 60$.

Finally, once the optimal classifier on the ROC convex hull generated on the training set is selected, we evaluate the generalization capability on the test set. Table 9 shows for all the classifiers the values of TPR, FPR, complexity, number of rules and cost, in the form (average ± the standard deviation), obtained repeating the selection process for each fold. Obviously, for the Chi FRBC and C4.5, we performed no selection, but we simply adopted the unique classifier generated by the learning process.

We can observe that, on average, the classifiers selected from the ROC convex hull generated by the three-objective evolutionary approach outperform in terms of cost all the other classifiers, while maintaining a low level of complexity (7.75 rules and 23.75 conditions on average). The classifiers generated by the two-objective evolutionary approaches executed on the original dataset are characterized by low level of complexity, but quite high costs. When we rebalance the datasets, the costs of these classifiers decrease, but the complexity increases. Thus, we can conclude that for this dataset the rebalancing helps the evolutionary process to explore more accurate, but also more complex solutions. As expected, the Chi FRBC is characterized by a very high level of complexity and also by a quite high cost. The ldc, naïve and C4.5 classifiers achieve quite low costs (almost comparable with the three-objective evolutionary approach). On the other hand, ldc and naïve are not interpretable at all since they cannot be expressed as sets of rules. As regards C4.5, a set of rules can be extracted from the decision tree produced by executing the rule learning algorithm based on the classical C4.5 decision tree algorithm (see Quinlan 1993 for a complete description of the algorithm). For this reason, in Table 9 we also show the average complexity and the number of rules for this classifier type. We can observe that the number of rules in C4.5 is higher than the number of rules in the FRBC selected by using our approach. Further, the fuzzy rules used in the FRBC are certainly more interpretable than the rules used in C4.5. As an example, Table 10 and Fig. 6 show, respectively, the rule base of an FRBC and some rules extracted from the C4.5 decision tree for one of the five folds. We can note that the fuzzy rules are easily comprehensible and provide very useful information to characterize a nodule. Indeed, we can learn that nodules are characterized by both a high value of sphericity and a low average gray level. On the other hand, the rules extracted from the C4.5 decision tree cannot be easily expressed in linguistic terms, since the crisp intervals used in the conditions do not correspond to intuitive partitions of the system variables. Thus, these rules result to be quite unreadable.

5.3.2 *Mammography dataset*

The Mammography dataset was used in Woods et al. (1993). It consists of 11,183 samples with 6 numeric features, and 2 classes representing, respectively, calcification (cancerous samples) and non calcification (non-cancerous samples). The minority class, which represents calcification, contains only 260 examples, i.e., 2.32% of the total examples. In this work, we used a filtered version of the dataset composed of 7,853 patterns: indeed, we carefully analyzed the dataset and eliminated all the patterns with zero value for all the features.

As an example of classifier selection, also for this dataset we fixed the a-priori probabilities $P(n)$ and $P(p)$ of the negative and positive classes to the class frequencies in the data set, that is, $P(n)$ equal to 7615/7853 ($\cong$97%) and $P(p)$ equal to 238/7853 ($\cong$3%). Then, we chose a ratio $c(Y, \mathrm{n})/c(N, \mathrm{p})$ equal to 1/60. As for the first dataset, this ratio was motivated by considering the cost of a false negative much higher than the cost of a false positive. We obtained a slope for the optimal iso-cost lines equal to 0.5.

Table 11 shows for each classifier the values of TPR, FPR, complexity, number of rules and cost of the best classifiers, in the form (average ± the standard deviation).

We can observe that, on average, the classifiers selected from the ROC convex hull generated by the three-objective evolutionary approach outperform in terms of cost all the

Fig. 5 Pareto front approximations and ROC convex hull on the training and test sets

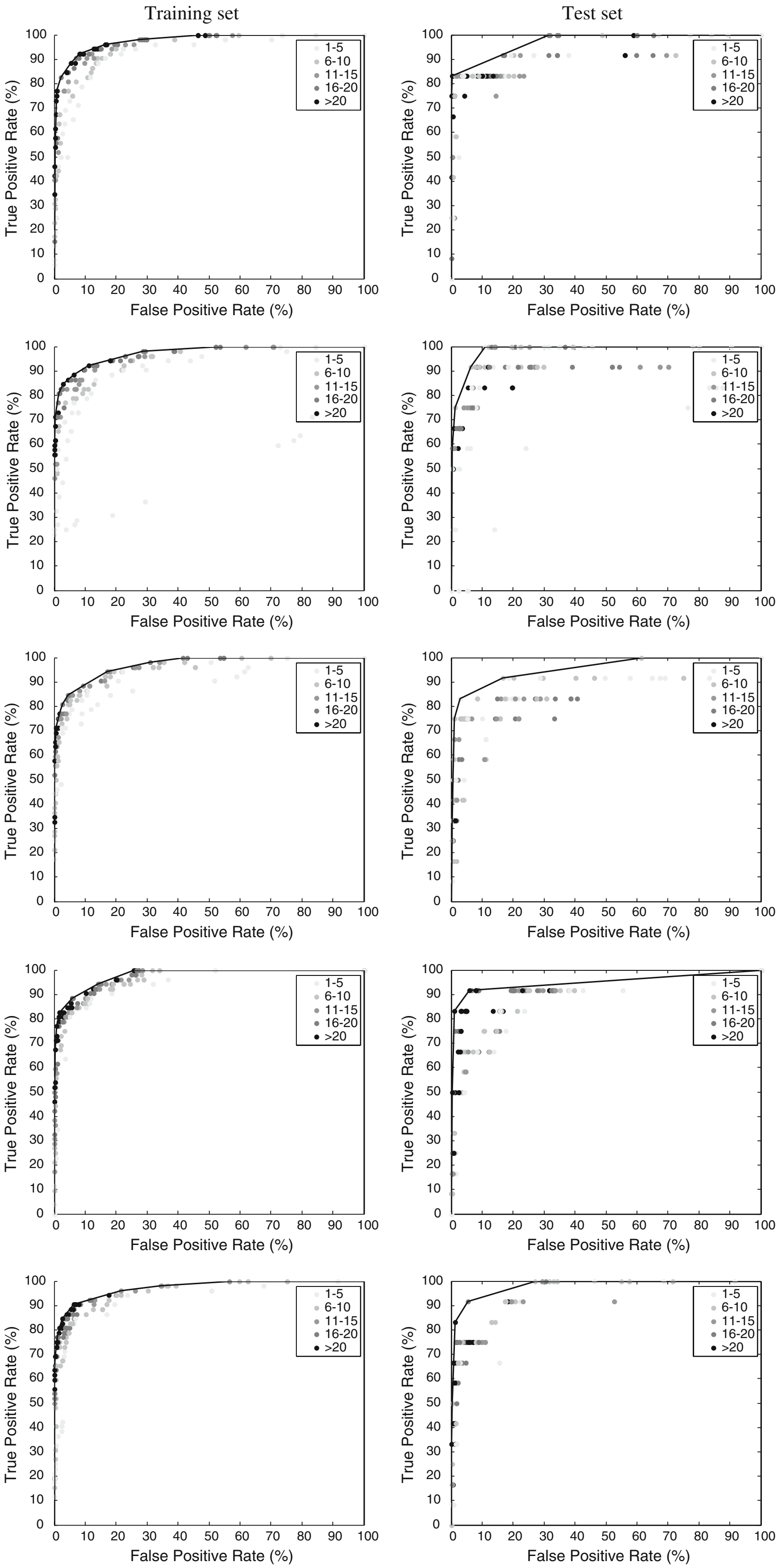

Table 9 FRBC selection: Results of the 5-fold cross-validation on Nodules dataset (test set)

Algorithm	TPR	FPR	Complexity	Rules	Cost
3ob_FRBC	87.50 ± 4.81	13.85 ± 5.02	23.75 ± 7.89	7.75 ± 2.63	0.35 ± 0.08
2ob_ FRBC_AUC	73.33 ± 23.86	5.67 ± 1.35	15.60 ± 6.62	5.00 ± 2.12	0.53 ± 0.42
2ob_FRBC_GM	70.00 ± 7.45	7.19 ± 2.88	27.00 ± 9.03	8.40 ± 2.88	0.61 ± 0.15
2ob_FRBC_AUC_Sm	76.67 ± 6.97	8.45 ± 3.41	37.80 ± 8.32	11.40 ± 2.70	0.50 ± 0.11
2ob_FRBC_GM_Sm	85.00 ± 6.97	9.82 ± 1.81	30.00 ± 15.25	9.00 ± 4.69	0.37 ± 0.13
Chi FRBC	75.00 ± 8.33	10.60 ± 2.96	411.20 ± 23.73	102.80 ± 5.93	0.55 ± 0.16
ldc	86.67 ± 7.45	15.75 ± 2.67			0.39 ± 0.15
Naïve	85.00 ± 6.97	13.75 ± 3.47			0.40 ± 0.14
C4.5	83.33 ± 5.89	7.72 ± 1.62	82.4 ± 5.12	28 ± 3.32	0.37 ± 0.11

Table 10 Rule base of an FRBC selected in one of five folds

Sphericity	Average gray level	Standard deviation of average gray level	Inverse of elongation	Class
HIGH	LOW	MEDIUM	MEDIUM LOW	Nodule
MEDIUM HIGH	LOW	LOW	–	Nodule
MEDIUM LOW	MEDIUM LOW	MEDIUM HIGH	MEDIUM HIGH	Non nodule
LOW	MEDIUM HIGH	MEDIUM LOW	MEDIUM	Non Nodule
–	–	HIGH	LOW	Non nodule
MEDIUM LOW	MEDIUM LOW	MEDIUM HIGH	LOW	Non nodule

IF (Std. Dev. Av. Gray Lev.<=0.5356) **THEN** Class=Non nodule

ELSE IF (Inverse of elongation>0.3333 && Sphericity>0.5 && Average gray level<=0.0497) **THEN** Class=Non nodule

ELSE IF (Sphericity>0.7 && Average gray level>0.0571 && Sphericity<=0.7129) **THEN** Class=Non nodule

ELSE IF (Sphericity>0.7 && Average gray level<=0.0571 && Inverse of elongation>0.1808 && Std. Dev. Av. Gray Lev.<=0.7262) **THEN** Class=Non nodule

ELSE IF (Inverse of elongation>0.3333 && Std. Dev. Av. Gray Lev.<=0.6537 && Sphericity>0.5 && Average gray level<=0.0497 && Average gray level>0.0301) **THEN** Class=Non nodule

ELSE IF (Std. Dev. Av. Gray Lev.>0.5484 && Std. Dev. Av. Gray Lev.<=0.6537 &&Inverse of elongation>0.3826 && Sphericity<=0.2087 && Sphericity>0.0062) **THEN** Class=Non nodule

…

ELSE IF (Std. Dev. Av. Gray Lev.>0.5484 && Sphericity<=0.5 && Sphericity>0.2087) **THEN** Class=Nodule

ELSE IF (Sphericity<=0.3711 && Std. Dev. Av. Gray Lev.>0.5439) **THEN** Class=Nodule

ELSE IF (Inverse of elongation<=0.4369 && Sphericity>0.7258 && Average gray level>0.0801 && Sphericity<=0.8191) **THEN** Class=Nodule

ELSE IF (Average gray level>0.0571 && Inverse of elongation<=0.3202) **THEN** Class=Nodule

…

ELSE Class=Non nodule

Fig. 6 Example of rules extracted from the C4.5 decision tree for Nodules dataset

other classifiers, though their complexities are slightly higher than the FRBCs generated by the two-objective evolutionary approaches. As regards these FRBCs, we can observe that the rebalancing does not affect particularly their performance. Indeed, both the cost and the complexity remain practically unchanged. The Chi FRBC is characterized by low cost, despite the highest level of complexity. As regards the three classifiers used for comparison, only C4.5 achieves costs comparable to the three-objective evolutionary approach, with a very low level of complexity. We recall, however, that the C4.5 rules are not easy to interpret.

Table 11 FRBC selection: Results of the fivefold cross-validation on mammography dataset (test set)

Algorithm	TPR	FPR	Complexity	Rules	Cost
3ob_FRBC	89.36 ± 7.16	14.26 ± 6.08	85.00 ± 15.60	16.75 ± 3.40	0.33 ± 0.13
2ob_ FRBC_AUC	84.26 ± 5.34	9.34 ± 3.06	67.00 ± 14.27	14.00 ± 2.83	0.37 ± 0.10
2ob_FRBC_GM	86.38 ± 8.46	11.57 ± 1.63	65.00 ± 18.34	13.20 ± 4.09	0.36 ± 0.16
2ob_FRBC_AUC_Sm	88.09 ± 7.15	18.03 ± 7.24	61.80 ± 35.52	12.80 ± 7.16	0.39 ± 0.14
2ob_FRBC_GM_Sm	86.81 ± 5.08	14.35 ± 5.18	64.60 ± 14.05	12.80 ± 2.68	0.38 ± 0.13
Chi FRBC	91.62 ± 3.36	23.38 ± 5.61	2073.00 ± 31.29	327.00 ± 10.34	0.35 ± 0.11
ldc	87.23 ± 8.90	27.67 ± 6.72			0.50 ± 0.15
Naïve	87.23 ± 4.26	27.43 ± 4.72			0.50 ± 0.10
C4.5	87.66 ± 8.95	14.84 ± 3.11	57.60 ± 9.72	17.40 ± 5.40	0.37 ± 0.19

6 Conclusions

In this paper, we have introduced a novel approach to determine FRBCs for highly imbalanced datasets. First, we have used the well-known NSGA-II multi-objective optimization algorithm to generate an approximation of a Pareto front composed of FRBCs with different trade-offs among sensitivity, specificity, and interpretability. Sensitivity and specificity are evaluated in the ROC plane, where each FRBC is described as a point whose coordinates are the true positive rate and the false positive rate, respectively. Then, after projecting the overall Pareto front onto the ROC plane, we have determined the potentially optimal classifiers by using the ROC convex hull method. To compare our approach with two evolutionary approaches, three well-known classifiers, namely, ldc, naïve, and C4.5, and one heuristic approach to FRBC generation, we have used 13 imbalanced datasets and the AUCH as metric. The Wilcoxon signed-rank test applied between our three-objective evolutionary approach and the other classifiers shows that our approach statistically outperforms the other approaches (except for ldc) in terms of AUCH, though the generated FRBCs are characterized by low level of complexity. Finally, we have shown a procedure to select optimal FRBCs among those in the convex hulls and have discussed the performance of these FRBCs for discriminating regions of interest (ROIs) in the lungs as nodules and non-nodules in a Computer Aided Diagnosis system, and for discriminating calcification (cancerous samples) and non calcification (non-cancerous samples) in mammography. On both these medical applications, the classification cost achieved by our approach has been lower than the ones obtained by the methods used for comparison.

References

Alcalá R, Gacto MJ, Herrera F, Alcalá-Fdez J (2007) A multi-objective genetic algorithm for tuning and rule selection to obtain accurate and compact linguistic fuzzy rule-based systems. Int J Uncertain Fuzziness Knowl Based Syst 15(5):521–537. doi:10.1142/S0218488507004856

Alcalá-Fdez J, Sánchez L, García S, del Jesus MJ, Ventura S, Garrell JM, Otero J, Romero C, Bacardit J, Rivas VM, Fernández JC, Herrera F (2009) KEEL: a software tool to assess evolutionary algorithms to data mining problems. Soft Comput 13(3):307–318. doi:10.1007/s00500-008-0323-y

Anastasio M, Kupinski M, Nishikawa R (1998) Optimization and FROC analysis of rule-based detection schemes using a multi-objective approach. IEEE Trans Med Imaging 17(10):1089–1093. doi:10.1109/42.746726

Antonelli M, Frosini G, Lazzerini B, Marcelloni F (2006) A CAD system for lung nodule detection based on an anatomical model and a fuzzy neural network. In: Proceedings of NAFIPS, Montreal, Canada, 3–6 June, pp 448–453

Asuncion A, Newman D (2007) UCI machine learning repository. University of California, Irvine, School of Information and Computer Sciences. http://www.ics.uci.edu/~mlearn/MLRepository.html

Awai K, Murao K, Ozawa A, Komi M, Hayakawa H, Hori S, Nishimura Y (2004) Pulmonary nodules at chest CT: effect of computer-aided diagnosis on radiologists' detection performance. Radiology 230(2):347–352. doi:10.1148/radiol.2302030049

Batista G, Prati R, Monard M (2004) A study of the behaviour of several methods for balancing machine learning. SIGKDD Explor 6(1):20–29. doi:10.1145/1007730.1007735

Casillas J, Cordon O, Del Jesus MJ, Herrera F (2001) Genetic feature selection in a fuzzy rule-based classification system learning process for high-dimensional problems. Inf Sci 136:135–157. doi:10.1016/S0020-0255(01)00147-5

Casillas J, Cordon O, Herrera F, Magdalena L (eds) (2003a) Accuracy improvements in linguistic fuzzy modeling. Springer, Berlin

Casillas J, Cordon O, Herrera F, Magdalena L (eds) (2003b) Interpretability issues in fuzzy modeling. Springer, Berlin

Casillas J, Cordon O, Del Jesus MJ, Herrera F (2005) Genetic tuning of fuzzy rule deep structures preserving interpretability and its interaction with fuzzy rule set reduction. IEEE Trans Fuzzy Syst 13(1):13–29. doi:10.1109/TFUZZ.2004.839670

Casillas J, Herrera F, Péreza R, Del Jesus MJ, Villar P (2007) Special issue on genetic fuzzy systems and the Interpretability-Accuracy Trade-off. Int J Approx Reason 44(1):1–3. doi:10.1016/j.ijar.2006.06.002

Chang X, Lilly JH (2004) Evolutionary design of a fuzzy classifier from data. IEEE Trans Syst Man Cybern B 34(4):1894–1906. doi:10.1109/TSMCB.2004.831160

Chawla N, Bowyer K, Hall L, Kegelmeyer W (2002) Smote: synthetic minority over-sampling technique. J Artif Intell Res 16:321–357

Chi Z, Yan H, Pham T (1996) Fuzzy algorithms with applications to image processing and pattern recognition. World Scientific, Singapore

Cococcioni M, Ducange P, Lazzerini B, Marcelloni F (2007) A Pareto-based multi-objective evolutionary approach to the identification of Mamdani fuzzy systems. Soft Comput 11(11):1013–1031. doi:10.1007/s00500-007-0150-6

Coello Coello CA, Lamont GB (2004) Applications of multi-objective evolutionary algorithms. World Scientific, Singapore

Coello Coello CA (2006) Evolutionary multi-objective optimization: a historical view of the field. IEEE Comput Intell Mag 1(1):28–36. doi:10.1109/MCI.2006.1597059

Cordon O, Del Jesus MJ, Herrera F (1999) A proposal on reasoning methods in fuzzy rule-based classification systems. Int J Approx Reason 20(1):21–45

Cordon O, Herrera F, Hoffmann F, Magdalena L (2001) Genetic fuzzy systems. World Scientific, Singapore

Cordon O, Del Jesus MJ, Herrera F, Magdalena L, Villar P (2003) A multiobjective genetic learning process for joint feature selection and granularity and contexts learning in fuzzy rule-based classification systems. In: Casillas J, Cordon O, Herrera F, Magdalena L (eds) Accuracy improvements in linguistic fuzzy modeling. Springer, Berlin, pp 79–99

Deb K (2001) Multi-objective optimization using evolutionary algorithms. Wiley, London

Deb K, Pratab A, Agarwal S, Meyarivan T (2002) A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE Trans Evol Comput 6(2):182–197. doi:10.1109/4235.996017

Duin RPW (2007) PRTools (Version 4.0) A Matlab toolbox for pattern recognition. Pattern recognition group, Delft, University of Technology. http://www.prtools.org

Everson RM, Fieldsend E (2006) Multiobjective optimization of safety related systems: An application short-term conflict alert. IEEE Trans Evol Comput 10(2):187–198. doi:10.1109/TEVC.2005.856067

Fawcett T (2003) ROC graphs: Notes and practical considerations for researchers. Tech. Rep. HPL-2003-4, HP Labs

Fawcett T (2006) An introduction to ROC analysis. Pattern Recognit Lett 27:861–874. doi:10.1016/j.patrec.2005.10.010

Fernandez A, García S, Del Jesus MJ, Herrera F (2008) A study of the behaviour of linguistic fuzzy rule based classification systems in the framework of imbalanced data sets. Fuzzy Sets Syst 159(18):2378–2398. doi:10.1016/j.fss.2007.12.023

Herrera F (2008) Genetic fuzzy systems: taxonomy, current research trends and prospects. Evol Intell 1:27–46. doi:10.1007/s12065-007-0001-5

Ho SY, Chen HM, Ho SJ, Chen TK (2004) Design of accurate classifiers with a compact fuzzy-rule base using an evolutionary scatter partition of feature space. IEEE Trans Syst Man Cybern 34(2):1031–1044. doi:10.1109/TSMCB.2003.819160

Horn J, Nafpliotis N, Goldberg DE (1999) A niched Pareto genetic algorithm for multiobjective optimization. In: Proceedings of the first IEEE conference on evolutionary computation, Orlando, Florida, 27–29 June, pp 82–87

Ishibuchi H (2007) Multiobjective genetic fuzzy systems: review and future research directions. In: Proceedings of the 2007 international conference on fuzzy systems, London, 23–26 July, pp 1-6

Ishibuchi H, Nojima Y (2007) Analysis of interpretability-accuracy tradeoff of fuzzy systems by multiobjective fuzzy genetics-based machine learning. Int J Approx Reason 44(1):4–31. doi:10.1016/j.ijar.2006.01.004

Ishibuchi H, Yamamoto T (2004) Fuzzy rule selection by multi-objective genetic local search algorithms and rule evaluation measures in data mining. Fuzzy Sets Syst 141(1):59–88. doi:10.1016/S0165-0114(03)00114-3

Ishibuchi H, Yamamoto T (2005) Rule weights specification in fuzzy rule-based classification systems. IEEE Trans Fuzzy Syst 13(4):428–435. doi:10.1109/TFUZZ.2004.841738

Ishibuchi H, Murata T, Turksen IB (1997) Single-objective and two-objective genetic algorithms for selecting linguistic rules for pattern classification problems. Fuzzy Sets Syst 89(2):135–150. doi:10.1016/S0165-0114(96)00098-X

Ishibuchi H, Nakashima T, Nii M (2005a) Classification and modeling with linguistic information granules: advanced approaches to linguistic data Mining. Springer, Berlin

Ishibuchi H, Nozaki K, Yamamoto N, Tanaka H (2005b) Selecting fuzzy if-then rules for classification problems using genetic algorithms. IEEE Trans Fuzzy Syst 3(3):260–270. doi:10.1109/91.413232

Karr CL, Gentry EJ (1993) Fuzzy control of pH using genetic algorithms. IEEE Trans Fuzzy Syst 1:46–53. doi:10.1109/TFUZZ.1993.390283

Knowles JD, Corne DW (2002) Approximating the non dominated front using the Pareto archived evolution strategy. Evol Comput 8(2):149–172. doi:10.1162/106365600568167

Kupinski M, Anastasio M (1999) Multiobjective genetic optimization of diagnostic classifiers with implications for generating receiver operating characteristic curves. IEEE Trans Med Imaging 18(8):675–685. doi:10.1109/42.796281

Mansoori E, Zolghadri M, Katebi S (2007) A weighting function for improving fuzzy classification systems performance. Fuzzy Sets Syst 158(5):583–591. doi:10.1016/j.fss.2006.10.004

Nakashima T, Shaefer G, Yokota Y, Ishibuchi H (2007) A weighted fuzzy classifier and its application to image processing tasks. Fuzzy Sets Syst 158:284–294. doi:10.1016/j.fss.2006.10.011

Nauck D, Kruse R (1999) Obtaining interpretable fuzzy classification rules from medical data. Artif Intell Med 16(2):149–169. doi:10.1016/S0933-3657(98)00070-0

Provost F, Fawcett T (2001) Robust classification for imprecise environments. Mach Learn 42(3):203–231. doi:10.1023/A:1007601015854

Quinlan JR (1993) C4.5: Programs for Machine Learning. Morgan Kauffman, San Mateo

Raskutti B, Kowalczyk A (2004) Extreme rebalancing for SVMs: a case study. SIGKDD Explor 6(1):60–69. doi:10.1145/1007730.1007739

Sheskin D (2003) Handbook of parametric and nonparametric statistical procedures. Chapman & Hall/CRC, London/Boca Raton

Srinivas N, Deb K (1998) Multi-objective function optimization using non-dominated sorting genetic algorithms. Evol Comput 2:221–248. doi:10.1162/evco.1994.2.3.221

Wilcoxon F (1945) Individual comparisons by ranking methods. Biometrics 1:80–83. doi:10.2307/3001968

Woods K, Doss C, Bowyer K, Solka J, Priebe J, Kegelmeyer P (1993) Comparative evaluation of pattern recognition techniques for detection of microcalcifications in mammography. Int J Pattern Recognit Artif Intell 7(6):1417–1436. doi:10.1142/S0218001493000698

Yen J, Wang L, Gillespie GW (1998) Improving the interpretability of TSK fuzzy models by combining global learning and local learning. IEEE Trans Fuzzy Syst 6(4):530–537. doi:10.1109/91.728447

Zitzler E, Thiele L (1999) Multiobjective evolutionary algorithms: a comparative case study and the strength Pareto approach. IEEE Trans Evol Comput 3:257–271. doi:10.1109/4235.797969

Zitzler E, Deb K, Thiele L (2000) Comparison of multiobjective evolutionary algorithms: empirical results. Evol Comput 8(2):173–195. doi:10.1162/106365600568202

Zitzler E, Laumanns M, Thiele L (2001) SPEA2: Improving the strength Pareto evolutionary algorithm for multiobjective optimization. In: Proceedings of EUROGEN2001 evolutionary methods for design, opt. and control with applications to industrial problems, Athens, pp 95–100