Comparison of Multi-objective Algorithms Applied to Feature Selection

Özlem Türkşen[1], Susana M. Vieira[2], José F.A. Madeira[2,3], Ayşen Apaydın[1], and João M.C. Sousa[2]

Abstract The feature selection problem can be formulated as a multi-objective optimization (MOO) problem, as it involves the minimization of the feature subset cardinality and the misclassification error. In this chapter, a comparison of MOO algorithms applied to feature selection is presented. The used MOO methods are: Nondominated Sorting Genetic Algorithm II (NSGA-II), Archived Multi Objective Simulated Annealing (AMOSA), and Direct Multi Search (DMS). To test the feature subset solutions, Takagi-Sugeno fuzzy models are used as classifiers. To solve the feature selection problem, AMOSA was adapted to deal with discrete optimization. The multi-objective methods are applied to four benchmark datasets used in the literature and the obtained results are compared and discussed.

1 Introduction

Generally, real-world data sets tend to be complex, very large, and normally contain many irrelevant features. One of the most important steps in data analysis for classification is feature selection, which has been an active research area on many fields, such as data mining, pattern recognition, image understanding, machine learning or statistics. The main idea of feature selection is to choose a subset of the available features, by eliminating redundant features with little or no predictive information. There are two key decisions involved in the feature subset selection problem: (i) the number of selected

[1] Ankara University, Faculty of Science, Statistics Department, 06100, Ankara, Turkey, {Ozlem.Turksen,Aysen.Apaydin}@science.ankara.edu.tr
[2] Technical University of Lisbon, Instituto Superior Técnico, Dept. of Mechanical Engineering, CIS - IDMEC/LAETA, Lisbon, Portugal, {susana.vieira,jmsousa}@ist.utl.pt
[3] ISEL, Rua Conselheiro Emidio Navarro, 1, 1959-007 Lisboa, Portugal, jaguilar@dem.ist.utl.pt

C. Borgelt *et al.* (Eds.) Towards Advanced Data Analysis, STUDFUZZ 285, pp. 359–375.
springerlink.com

features and (ii) the best features to be selected [28]. An effective feature selection method can minimize the classification error, improve the prediction accuracy, and also discover the relevant features. From this point of view, it is possible to say that feature subset selection is a multi-objective optimization problem. Recently, the multi-objective feature selection (MOFS) problem was addressed in many studies. In [14], the feature selection problem is seen as a multi-objective problem and the Niched Pareto Genetic Algorithm (NPGA), which uses a commonality-based crossover operator, is applied to solve the MOFS problem, using neural models as classifiers. A variation of NPGA with a one-nearest neighbor classifier is applied to the MOFS problem in [13]. In [15], the application of the Multi Objective Genetic Algorithm (MOGA) is proposed for feature subset selection on a number of neural and fuzzy models together with fast subset evaluation techniques. Further, MOGA was used with different classifiers, namely, fuzzy rule based classification [8], back-propagation neural networks [19], and support vector machines [5], to solve the MOFS problem. There are some studies on the use of the nondominated sorting genetic algorithm (NSGA), firstly proposed in [26], for MOFS with different wrapper methods, such as neural networks [23] and decision trees [32], on different data sets. In [17, 22, 30, 12, 18, 24], the nondominated sorting genetic algorithm II (NSGA-II), one of the most efficient multi-objective algorithms, was used to solve the MOFS problem using different classification methods.

Archived multi-objective simulated annealing (AMOSA), which was proposed in [3], is an efficient multi-objective version of the simulated annealing algorithm, based on Pareto dominance. AMOSA incorporates a novel concept of the amount of dominance, in order to determine the acceptance of a new solution, as in NSGA-II. AMOSA was mainly used for continuous multi-objective problems, except in [31] where it was applied to gene selection. In this chapter, AMOSA was adapted for the feature selection problem, and it is called Modified AMOSA.

Direct multisearch (DMS) is a novel MOO algorithm proposed in [9]. This method is inspired by the search/poll paradigm of direct-search methods of the directional type and uses the concept of Pareto dominance to maintain a list of non-dominated points (from which the new iterates or poll centers are chosen). The aim of this method is to generate as many points in the Pareto front as possible from the polling procedure itself, while keeping the whole framework general enough to accommodate other disseminating strategies. This chapter presents a comparison of derivative-free multi-objective algorithms, which are NSGA-II, Modified AMOSA and DMS, for the feature selection problem with two different objectives: minimizing the number of features and minimizing the misclassification rate. The chapter is organized as follows: the next section presents a multi-objective formulation of the feature selection problem and a brief description of fuzzy modeling for classification. The derivative-free multi-objective algorithms for MOFS, namely NSGA-II, Modified AMOSA and DMS, are presented in Section 3. In Section 4, the

results obtained for several benchmark databases are presented, and a comparison of the studied multi-objective algorithms is made. Some conclusions are drawn in Section 5 and possible future work is discussed.

2 Feature Selection

Feature selection is the process of selecting a subset of the available features to use in empirical modeling. A fundamental problem of feature selection is to determine a minimal subset of n features from the complete set of the features $\mathbf{x} = \{x_1, x_2, \ldots, x_N\}$, with $n < N$, without sacrificing accuracy. This means that there are 2^N possible feature subsets, which makes a brute-force approach (enumerating and testing all feature subsets) infeasible in most cases. It can be said that the main goal of feature selection is to reduce the number of features used in classification while maintaining an acceptable classification accuracy. From this perspective, a feature selection problem can be seen as a multi-objective problem with two objectives: minimization of the number of features and of the error rate of the classifier. In this chapter, a fuzzy classifier is built for each feature subset to evaluate the classification error. The solutions are evaluated by using fuzzy models, as they are universal approximators and can be interpretable under certain conditions.

2.1 Feature Selection as a Multi-objective Optimization Problem

A multi-objective optimization problem can be mathematically formulated as (see [21] for a more complete treatment):

$$\begin{aligned} &\min\ F(\mathbf{x}) \equiv (f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_m(\mathbf{x}))^T \\ &\textit{subject to} \quad \mathbf{x} \in \Omega \end{aligned} \tag{1}$$

where x is the vector of decisions or design variables belonging to the feasible region $\Omega \subseteq \mathbb{R}^n$ and $F(\mathbf{x}) \in \mathbb{R}^m$ is a vector of m objective functions. The word "min" in (1) means that we want to minimize all objective functions simultaneously. Note that this is a general formulation, which exploits that maximizing an objective function f_j is equivalent to minimizing $-f_j$.

The solution of the problem given in (1) is a set of solutions that are called Pareto optimal in the general framework of multi-objective optimization. A solution is said to be Pareto optimal if it is not dominated by any other solution available in the search space Ω (see [21] for a more complete treatment).

In feature selection it is common to encode a feature subset as a binary vector $\mathbf{x} \in \{0,1\}^N$, where N is the number of available features. This vector states for each feature whether it is selected ($x_i = 1$) or not ($x_i = 0$), see Fig. 1. The MOO algorithm for the feature selection problem consists of finding the set of features that simultaneously minimize two objectives: $f_1(\mathbf{x}) = \sum_{k=1}^{N} \mathbf{x}_k$ which is the number of selected features and $f_2(\mathbf{x}) = (1 - accuracy(\mathbf{x}))$ which is equivalent to maximize the accuracy.

2.2 Fuzzy Modeling for Feature Selection

Fuzzy models are suitable to deal with vague, imprecise and uncertain knowledge and data. These models use rules and logical connectives to establish relations between the features defined to derive the model. Three general methods for fuzzy classifier design can be distinguished [4, 25]: the regression method, the discriminant method and the maximum compatibility method. In the discriminant method, the classification is based on the largest discriminant function, which is associated with a certain class, regardless of the values or definitions of other discriminant functions. Hence, the classification decision does not change if the discriminant functions are transformed monotonically. Since this is a useful property, the discriminant method is used in this work for classification. The discriminant functions can be implemented as fuzzy inference systems, which can be Takagi-Sugeno (TS) fuzzy models [27]. When TS fuzzy systems are used, each discriminant function consists of rules of the type

$$\begin{aligned} \text{Rule } R_i^c : \text{If } x_1 \text{ is } A_{i1}^c \text{ and } \ldots \text{ and } x_n \text{ is } A_{in}^c \\ \text{then } d_i^c(\mathbf{x}) = f_i^c(\mathbf{x}),\ i = 1, 2, \ldots, K, \end{aligned}$$

where c is the number of classes, K is the number of fuzzy rules, and f_i^c is the consequent function for rule R_i^c. Please note that the antecedent parts of the rules can be different for different discriminants, as well as the consequents. Therefore, the output of each discriminant function $d_c(\mathbf{x})$ can be interpreted as a score (or evidence) for the associated class c given the input feature vector $\mathbf{x}_n$. The discriminant function for class c, with $c = 1, \ldots, C$ is computed by aggregating the contributions of the individual rules:

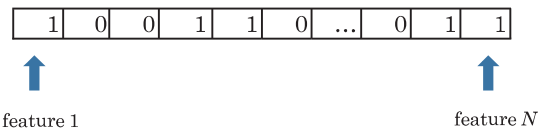

Fig. 1 Binary vector with N components.

$$d_c(\mathbf{x}) = \frac{\sum_{i=1}^{K} \beta_i f_i^c(\mathbf{x})}{\sum_{i=1}^{K} \beta_i}. \quad (2)$$

where $\beta_i = \prod_{j=1}^{n} \mu_{A_{ij}^c}(\mathbf{x})$ is the degree of activation of rule i of class c and $\mu_{A_{ij}^c}(\mathbf{x}) : \mathbb{R} \rightarrow [0, 1]$.

The input data consists of tuples $\mathbf{x}_i = [x_{i1}, x_{i2}, \ldots, x_{iM}, y_i]^T$with $i \in \{1, \ldots, N\}$, where the x_{ij}, $j \in 1, ...M$, are the values of the features and y_i is the class of the ith case. From this data and the number of rules K, the antecedent fuzzy sets A_{ij}, and the consequent parameters are determined by means of fuzzy clustering [25]. Note that the class is used as an input to the clustering algorithm. This paper uses the Gustafson-Kessel (GK) [16] clustering algorithm to compute the fuzzy partition matrix. Each identified cluster provides a local characteristic behavior of the system and each cluster defines a rule. The consequent parameters for each rule are obtained as a weighted ordinary least-square estimate.

3 Multi-objective Algorithms for Feature Selection

Derivative-free multi-objective algorithms can be used for feature selection problems, as most of these algorithms can deal with a set of possible solutions simultaneously. This means that several members of the Pareto optimal set can be found in a single run without any assumptions on continuity and differentiability of functions [23]. A comprehensive review of multi-objective algorithms can be found in [6, 10]. In this chapter, NSGA-II, Modified AMOSA, and DMS, which are described in the following sections, are used for the optimization of the MOFS problem.

3.1 NSGA-II for Multi-objective Feature Selection

NSGA-II has been successfully applied in many applications, such as image processing, bioinformatics, etc. [11]. The main characteristic of this algorithm is to use the fast non-dominated sorting technique and a crowding distance to construct population fronts that dominate each other in a domination rank. To implement NSGA-II in the MOFS problem, first a random population representing different points in the search space is created with the size n_{pop}, as in [18]. Each chromosome $C_i, i = 1, 2, \ldots, n_{pop}$, is binary, and the encoding of a chromosome was represented in Fig. 1, where a chromosome has n bits equal to "1". These features are the ones used to construct the fuzzy classifier. The search process of NSGA-II continues until the number of generations n_{gen} is reached. Tournament selection is used based on two criteria: rank and crowding distance, with rank taking precedence, see [11]. NSGA-II also

Table 1 Definition of the parameters for Modified AMOSA

Parameter	Description
Archive	Set of nondominated solutions
AL	Limit size of the Archive
β	Constant for the initial size of the Archive
T_{max}	Maximum (initial) temperature
T_{min}	Minimum (final) temperature
α	Cooling rate in simulated annealing
iter	Number of iterations at each temperature

incorporates an elitism scheme to maintain the best solutions; individuals with higher crowding distances have higher fitness values. NSGA-II for MOFS uses the single-point crossover method, which randomly chooses a crossing site along the string and exchanges all bits on the right side of the crossing site [10]. The mutation operator used in this work is the uniform mutation operator [20], which operates on each bit separately and randomly changes the bit's. If the chromosome is filled with "0"s, i.e., the feature subset is empty, a gene of the chromosome is selected randomly and replaced by "1" to obtain a non-empty feature subset.

3.2 Modified AMOSA for Multi-objective Feature Selection

AMOSA is a generalized version of simulated annealing for multi-objective optimization problems, which was proposed in [3]. AMOSA, a Pareto dominance based simulated annealing method, incorporates the concept of an *Archive* where the nondominated solutions seen so far are stored. In contrast to the original suggestion [3], we use a fixed-sized *Archive*, AL, instead of a soft and a hard limit. Similarly to the original AMOSA, initially a random *Archive* is generated with $\beta \times AL$ solutions, where β is a constant, $0 < \beta \leq 3$. As in NSGA-II, the solutions are encoded as a binary vector. The fitness evaluation is done for each *Archive* solution. The solutions in the *Archive* are sorted by using a domination relation. Only the obtained nondominated solutions are used to initialize the *Archive*. Here, the fast nondominated sorting mechanism is used to rank the solutions [11]. During the search, if the number of solutions in the *Archive* is higher than AL, the first AL solutions in the rank order are chosen from the *Archive*. Eventually, the ranked "1" solutions will constitute the Pareto front. Table 1 describes the parameters that need to be set *a priori*.

The search is started with a solution that is chosen randomly from the solutions in the *Archive*. This is taken as the current feature subset $F_{current}$, or

Algorithm 1 Modified AMOSA for feature selection

```
Set AL, β, T_max,T_min, α, iter and T = T_max.
Initialize the Archive.
Choose randomly a subset of features from the nondominated solutions in the Archive
and set it as F_current
Compute the probability of being accepted P_current using (4).
while T > T_min do
  for k = 0 to iter do
    Create a new subset of features F_new, using the perturbation scheme
    Compute P_new
    Check the domination status of F_new with respect to the F_current and the present
    solutions in the Archive
    Select the next F_current using the domination status of the original AMOSA [3].
  end for
  update T = α × T
  if | Archive |≥ AL then
    choose the first AL solutions from the Archive.
  end if
end while
```

the initial solution, at temperature $T = T_{max}$. To create a new solution, the $F_{current}$ solution is perturbed and a new feature subset F_{new} is generated and the solution is evaluated using the objective function. The perturbation is computed using a Laplace distribution to determine which decision variables should be mutated. Thereafter, the domination status of F_{new} is checked with respect to $F_{current}$ and to the solutions in the *Archive*. The nondominated solutions are defined using the acceptance concept from the original AMOSA [3]. Given two solutions a and b, where a dominates b, the amount of domination is defined as follows:

$$\Delta dom_{a,b} = \prod_{i=1, f_i(a) \neq f_i(b)}^{m} \left(\frac{| f_i(a) - f_i(b) |}{R_i} \right) \tag{3}$$

where m is the number of objectives, $f_i(a)$ and $f_i(b)$ are the ith objective values for the two different solutions and R_i is the range of the objective function. R_i is determined using the solutions in the *Archive*, in the current and in the new feature subsets. Based on the domination status, a number of cases can arise: (i) accept F_{new}, (ii) accept $F_{current}$ or (iii) accept a solution from the *Archive*. The acceptance is calculated based on the applicable case. The *Archive* limit is maintained and the content is continuously updated during the search. Whenever an unfavorable move is considered for acceptance, the probability of acceptance is calculated as:

$$P = \frac{1}{1 + \exp(\Delta dom \times T)} \tag{4}$$

where T is the temperature, and Δdom is calculated as in (3) [31]. The process is repeated *iter* times for each temperature, which is annealed with a cooling rate of $\alpha < 1$ until the minimum temperature T_{min} is reached. The process then stops, and the *Archive* contains the nondominated solutions. The steps of the Modified AMOSA applied to the feature selection problem are presented in Algorithm 1.

3.3 DMS for Multi-objective Feature Selection

Direct MultiSearch (DMS) is a novel derivative-free method for multi-objective optimization, which does not aggregate any of the objective functions [9]. DMS extends to MOO all types of direct-search methods that are of a directional type such as pattern search and generalized pattern search, generating set search, and mesh adaptive direct search [7]. This approach is called direct multisearch since it naturally generalizes direct search (of directional type) from single to multi-objective optimization.

The principles of DMS are extremely simple. Instead of updating a single point per iteration, it updates a list of feasible nondominated points. For a more detailed description of the algorithm please see [9]. The original DMS was developed for real-valued variables in the search space. This algorithm was adapted to cope with the discrete feature selection optimization problem. Each DMS solution, $\mathbf{x}_{DMS} = (x_1, \ldots, x_N) \in \mathbb{R}^N$ is converted into a binary solution, $\mathbf{x} = (x_1, \ldots, x_N), \forall i = 1, \ldots, N : x_i \in \{0, 1\}$, using the threshold

$$\delta = 0.5, \text{ i.e., } x_i = \begin{cases} 1 \; if \; x_i \geq 0.5 \\ 0 \; if \; x_i < 0.5 \end{cases} \tag{5}$$

4 Experimental Results

The derivative-free multi-objective algorithms for feature selection are applied to data sets taken from some well known benchmarks in the UCI Machine Learning Repository [2].

4.1 Data Sets

Wisconsin breast cancer original (WBCO), Wisconsin diagnostic breast cancer (WDBC), Wisconsin prognostic breast cancer (WPBC) and Sonar, were used to test the NSGA-II, Modified AMOSA, and DMS algorithms. Table 2 summarizes some general information regarding these datasets.

Table 2 Description of the used data sets.

No	data sets	# features	Classes	# samples
1	WBCO	9 (integer)	2	699
2	WDBC	32 (real)	2	569
3	WPBC	34 (real)	2	198
4	Sonar	60 (real & integer)	2	208

The MOO algorithms were implemented in Matlab. The parameter settings of NSGA-II and modified AMOSA used in the experiments are presented in Table 3 and Table 4, respectively.

Table 3 NSGA-II parameter values used in the experiments.

Data set	n_{pop}	P_{cross}	P_{mut}	n_{gen}
WBCO	100	0.8	1/9	100
WDBC	100	0.8	1/30	200
WPBC	100	0.8	1/32	200
Sonar	100	0.8	1/60	500

Table 4 Modified AMOSA parameter values used in the experiments.

Data set	AL	β	T_{max}	T_{min}	α	$iter$
WBCO	100	1.5	200	0.001	0.81	50
WDBC	100	1.5	200	0.001	0.81	400
WPBC	100	1.5	200	0.001	0.81	400
Sonar	100	1.5	200	0.001	0.81	500

The default parameters of DMS were used (version 0.2, May 2011) without cache. This DMS version is freely available for research, educational or commercial use, under a GNU lesser general public license [1].

In this chapter, 10-fold cross validation is used. After the application of multi-objective algorithms and selection of the features, the fuzzy classification models are validated using the test subsets. The prediction performance of the classifier is estimated by considering the average classification accuracy of the 10-fold cross validation experiments.

4.2 Results for Different Data Sets

The NSGA-II, Modified AMOSA, and DMS methods are applied to minimize both the size of feature subsets and the average misclassification rates for all data sets. The number of fuzzy rules is equal to 2 for WBCO and 3 for the other databases used in this work.

4.2.1 Wisconsin Breast Cancer Original

The WBCO data is widely used to test the effectiveness of classification algorithms. The aim of the classification is to distinguish between benign and malignant cancers based on nine measurements (attributes): clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli and mitoses. The attributes have integer values in the range $[1, 10]$. The original database contains 699 instances. However 16 of these are excluded as they are incomplete, which is common in data mining. The class distribution is 65.5% benign and 34.5% malignant [29].

Table 5 Feature subsets for WBCO data set with 10 fold cross-validation.

NF	Selected features			Value of 1-accuracy (%)
	NSGA-II	Modified AMOSA	DMS	
1	$\{2\}$	$\{2\}$	$\{2\}$	7.153
2	$\{2, 6\}$	$\{2, 6\}$	$\{2, 6\}$	4.721
3	$\{1, 2, 6\}$	$\{1, 2, 6\}$	-	4.435
3	-	-	$\{1, 3, 6\}$	3.720
4	-	-	$\{1, 3, 4, 6\}$	3.577
5	-	-	$\{1, 2, 3, 5, 6\}$	3.434

Table 5 shows the set of nondominated solutions obtained by NSGA-II, Modified AMOSA, and DMS. The obtained feature subsets are similar for the three algorithms.

In Figure 2, the nondominated solutions for each algorithm are presented, and it is shown that DMS presents the best results.

4.2.2 Wisconsin Diagnostic Breast Cancer

In WDBC, features are computed from a digitized image of a fine needle aspirate of a breast mass. They describe characteristics of the cell nuclei present in the image. The attribute information for WDBC is as follows: ID numbers,

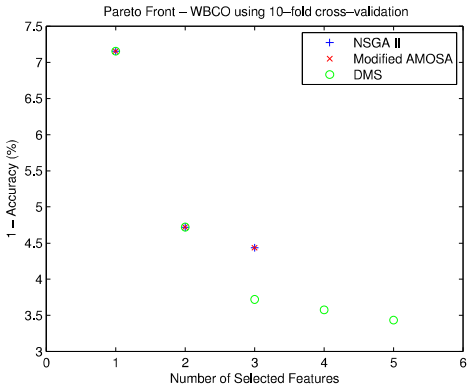

Fig. 2 Comparison of algorithms for WBCO data set.

diagnosis (M = malignant, B = benign) and ten real-valued features are computed for each cell: nucleus radius (mean of distances from center to points on the perimeter), texture (standard deviation of gray-scale values), perimeter, area, smoothness (local variation in radius lengths), compactness, concavity (severity of concave portions of the contour), concave points (number of concave portions of the contour), symmetry and fractal dimension [2].

Table 6 Feature subsets for WDBC data set with 10 fold cross validation.

NF	Selected features			Value of 1-accuracy (%)
	NSGA-II	Modified AMOSA	DMS	
1	{24}	-	{24}	8.260
2	{24, 28}	-	-	6.151
2	-	-	{21, 25}	4.745
3	{22, 24, 28}	-	-	4.569
3	-	-	{2, 21, 25}	4.042
3	-	-	{2, 24, 28}	4.042
3	-	-	{22, 24, 29}	4.042
3	-	-	{24, 25, 29}	4.042
4	{8, 22, 24, 25}	-	-	4.394
4	-	{21, 26, 29, 30}	-	6.503
4	-	-	{2, 3, 24, 25}	3.339
4	-	-	{2, 21, 28, 29}	3.339
5	{8, 10, 22, 24, 25}	-	-	3.866
5	-	{2, 3, 8, 21, 29}	-	5.800
5	-	-	{2, 14, 21, 25, 28}	2.812
5	-	-	{14, 21, 22, 25, 28}	2.812
6	-	-	{2, 14, 21, 25, 28, 29}	2.636
6	-	-	{2, 14, 24, 25, 28, 29}	2.636

The obtained nondominated solutions using NSGA-II, Modified AMOSA and DMS are summarized in Table 6. Figure 3 presents the average misclassification rates and feature subset cardinality of the three algorithms. DMS is clearly the best, followed by NSGA-II and Modified AMOSA, which yields quite poor results when compared to the other two algorithms.

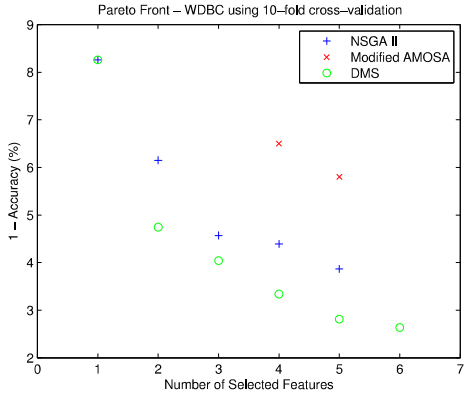

Fig. 3 Comparison of algorithms for WDBC data set.

4.2.3 Wisconsin Prognostic Breast Cancer

In this dataset, each record represents follow-up data for one breast cancer case. These are consecutive patients and include only those cases exhibiting invasive breast cancer and no evidence of distant metastases at the time of diagnosis. The first 30 features of the WPBC data set are computed from a digitized image of a fine needle aspirate of a breast mass. They describe characteristics of the cell nuclei present in the image [2]. Table 7 shows the feature subsets obtained by DMS, NSGA-II and Modified AMOSA. Figure 4 presents the results obtained by the three algorithms. DMS is the best of the three algorithms, and can find much more points on the Pareto front. In this case, NSGA-II and Modified AMOSA present similar results, both yielding a small number of solutions.

4.2.4 Sonar Data Set

The sonar data set contains information of 208 objects and 60 attributes. The objects are classified in two classes: “rock” and “mine”. A data frame with 208 observations and 61 variables is used. The first 60 represent the energy

Table 7 Feature subsets for WPBC data set with 10 fold cross validation.

NF	Selected features			Value of 1-accuracy (%)
	NSGA-II	Modified AMOSA	DMS	
1	{25}	-	-	23.74
1	-	-	{5}	22.73
2	{1, 25}	-	{1, 25}	19.70
2	-	-	{1, 22}	19.70
3	-	{11, 23, 24}	-	21.72
3	-	-	{1, 13, 25}	18.18
4	-	{1, 3, 7, 22}	-	18.69
4	-	-	{1, 13, 22, 32}	17.17
5	{1, 6, 8, 13, 25}	-	-	19.19
5	-	-	{1, 13, 20, 22, 32}	16.67
5	-	-	{1, 13, 24, 26, 32}	16.67
6	{1, 6, 8, 13, 19, 25}	-	-	18.18
6	-	-	{1, 6, 11, 13, 18, 32}	16.16
6	-	-	{1, 13, 22, 26, 27, 32}	16.16
7	-	-	{1, 13, 20, 22, 26, 27, 32}	15.66
10	-	{1, 2, 5, 8, 13, 14, 15, 17, 22, 24}	-	18.18
12	-	-	{1, 2, 6, 11, 12, 13, 14, 17, 18, 22, 24, 32}	14.65
14	-	-	{1, 2, 7, 9, 12, 13, 14, 17, 18, 20, 22, 24, 26, 29}	13.64
20	-	-	{1, 2, 5, 9, 12, 13, 14, 16, 17, 18, 20, 21, 22, 24, 25, 26, 27, 28, 29, 31}	13.13

within a particular frequency band, integrated over a certain period of time. The last column contains the class labels. There are two classes "0" if the object is a rock, and "1" if the object is a mine (metal cylinder) [2].

This data set is an interesting challenge for the proposed algorithm as the number of features is bigger than the usual benchmark examples. The obtained feature subsets are given in Table 8. Note that Modified AMOSA cannot find models with less than 13 features. On the other hand, NSGA-II has only results up to 7 features.

Figure 5 show the results obtained by the three algorithms. DMS presents clearly the best results, also for a wider spread of number of features. NSGA-II has good results for a small number of features. The results using Modified AMOSA are far from the Pareto front.

4.2.5 Discussion

By analyzing and comparing the four datasets, it can be concluded that Modified AMOSA was not able to deal with the feature cardinality in the

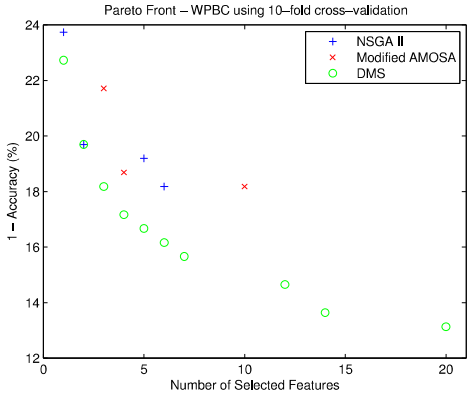

Fig. 4 Comparison of algorithms for WPBC data set.

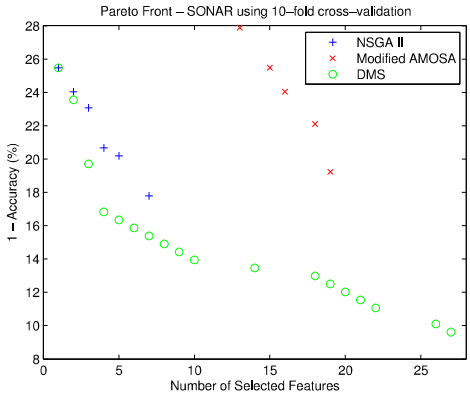

Fig. 5 Comparison of algorithms for the sonar data set.

objective function. NSGA-II proved to be a good algorithm for MOO, as expected. However, the very recent DMS algorithm is clearly the best of the three tested algorithms.

5 Conclusions

In this chapter, the feature selection problem is approached as a multi-objective problem. Two of the most important objectives for the feature selection problem were addressed: the minimization of the feature subset cardinality and the minimization of the classification error.

Table 8 Feature subsets for sonar data set with 10 fold cross validation.

NF	Selected features			Value of 1-accuracy (%)
	NSGA-II	Modified AMOSA	DMS	
1	{11}	-	{11}	25.48
2	{11, 16}	-	-	24.04
2	-	-	{11, 17}	23.56
3	{11, 16, 21}	-	-	23.08
3	-	-	{11, 18, 19}	19.71
4	{11, 15, 21, 46}	-	-	20.67
4	-	-	{11, 17, 19, 52}	16.83
5	{11, 15, 21, 38, 46}	-	-	20.19
5	-	-	{11, 17, 36, 45, 54}	16.35
6	-	-	{3, 9, 11, 18, 19, 54}	15.87
7	{3, 4, 11, 14, 21, 36, 46}	-	-	17.79
7	-	-	{3, 9, 11, 18, 19, 54, 60}	15.38
8	-	-	{5, 11, 17, 18, 36, 39, 46, 59}	14.90
9	-	-	{5, 11, 14, 17, 18, 36, 39, 46, 59}	14.42
10	-	-	{5, 11, 14, 17, 18, 36, 39, 41, 46, 59}	13.94
13	-	{1, 4, 7, 12, 13, 15, 16, 17, 23, 27, 36, 49, 59}	-	27.88
14	-	-	{7, 11, 17, 18, 23, 25, 30, 31, 35, 36, 44, 49, 50, 52}	13.46
15	-	{3, 4, 11, 12, 13, 15, 17, 29, 30, 34, 35, 40, 49, 59, 60}	-	25.48
16	-	{1, 3, 4, 8, 9, 12, 13, 15, 24, 25, 27, 33, 35, 36, 40, 47}	-	24.04
18	-	{3, 11, 12, 13, 17, 21, 30, 34, 35, 36, 40, 45, 49, 50, 55, 57, 58, 60}	-	22.21
18	-	-	{4, 7, 11, 13, 17, 18, 24, 29, 30, 31, 32, 34, 35, 36, 47, 49, 50, 51}	12.98
19	-	{10, 11, 12, 15, 17, 18, 21, 25, 29, 30, 34, 35, 45, 47, 49, 50, 52, 59, 60}	-	19.23
19	-	-	{4, 7, 11, 13, 17, 24, 29, 30, 31, 32, 34, 35, 36, 47, 49, 50, 51, 55, 58}	12.50
20	-	-	{4, 7, 11, 13, 17, 24, 29, 30, 31, 32, 34, 35, 36, 46, 47, 49, 50, 51, 55, 58}	12.50
21	-	-	{8, 11, 17, 18, 20, 21, 24, 27, 30, 31, 32, 35, 36, 39, 40, 43, 49, 50, 53, 60}	11.54
22	-	-	{8, 11, 17, 18, 19, 20, 21, 24, 27, 30, 31, 32, 35, 36, 39, 40, 43, 48, 49, 50, 53, 60}	11.06
26	-	-	{4, 8, 11, 17, 18, 19, 20, 21, 24, 27, 29, 30, 31, 32, 34, 35, 36, 39, 40, 43, 45, 47, 49, 50, 53, 55}	10.10
27	-	-	{4, 8, 11, 17, 18, 19, 20, 21, 24, 27, 29, 30, 31, 32, 34, 35, 36, 39, 40, 43, 45, 47, 49, 50, 53, 55, 60}	9.615

Archived multi-objective simulated annealing was adapted to cope with the feature selection problem. The modified AMOSA is compared with two multi-objective optimization algorithms: NSGA-II and DMS. In order to evaluate the feature subsets, fuzzy models are used.

Both NSGA-II and DMS outperformed the proposed modified AMOSA. NSGA-II is a population based multi-objective algorithm, which showed to be more efficient in approximating the Pareto front. DMS also progresses with the evolution of a set of nondominated solutions, and the greedy properties of the search mechanism granted a better performance for approximating the Pareto front. One of the key mechanisms in modified AMOSA is the creation of a new solution by using the neighborhood of the current solution, which is called perturbation. The results showed that this search mechanism is not effective. Thus, the perturbation scheme should be improved.

Acknowledgements The research in this work has been supported by the COST Action IC0702 STSMs. The research by Özlem Türkşen was partially supported by the TUBITAK (The Scientific and Technological Research Council of Turkey-code 2214-Research Project) which is gratefully acknowledged. This work was also supported by ISEL, byFundação para a Ciência e a Tecnologia (FCT), through IDMEC-IST under LAETA, and by a FCT grant SFRH/BPD/65215/2009, Ministério do Ensino Superior, da Ciência e da Tecnologia, Portugal.

References

1. Direct multisearch (dms) for multi-objective optimization. http://www.mat.uc.pt/dms/ (2012)
2. Asuncion A, Newman D (2007) UCI machine learning repository. http://www.ics.uci.edu/∼mlearn/MLRepository.html
3. Bandyopadhyay S, Saha S, Maulik U, Deb K (2008) A simulated annealing-based multi-objective optimization algorithm: AMOSA. *IEEE Transactions on Evolutionary Computation* 12(3):269–283
4. van den Berg J, Kaymak U, van den Bergh WM (2002) Fuzzy classification using probability-based rule weighting. *Proc. IEEE Int. Conf. on Fuzzy Systems*, 2:991–996. IEEE Press, Piscataway
5. Bhatia S, Prakash P, Pillai G (2008) SVM based decision support system for heart disease classification with integer-coded genetic algorithm to select critical features. *Proc. World Congress on Engineering and Computer Science*. San Francisco
6. Coello CC, Voldhuizen D, Lament G (eds.) (2002) *Evolutionary Algorithms for Solving Multi Objective Problems*. Kluwer Academic, New York
7. Conn AR, Scheinberg K, Vicente LN (2009) Introduction to derivative-free optimization. *MPS-SIAM Series on Optimization*. SIAM, Philadelphia
8. Cordon O, Herrera F, Jesus M, Magdalena L, Sanchez A, Villar P (2002) A multiobjective genetic algorithm for feature selection and data base learning in fuzzy-rule based classification systems. *Proc. 9th Int. Conf. on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU 2002)*, 823–830. Annecy, France
9. Custódio AL, Madeira JFA, Vaz AIF, Vicente LN (2011) Direct multisearch for multiobjective optimization. *SIAM Journal on Optimization* 10–18
10. Deb K, ed. (2004) Multi Objective Optimization using Evolutionary Algorithms. J. Wiley & Sons, Chichester
11. Deb K, Pratap A, Agarwal S, Meyarivan T (2002) A fast and elitist multi-objective genetic algorithm: NSGA-II. *IEEE Trans. an Evolutionary Computation* 6(2):10–18
12. Ekbal A, Saha S, Garbe C (2010) Feature selection using multi-objective optimization for named entity recognition. *Proc. IEEE Int. Conf. on Pattern Recognition*, 1937–1940
13. Emmanouilidis C, ed. (2002) Evolutionary Multi-Objective Feature Selection and ROC Analysis with Application to Industrial Machinery Fault Diagnosis. *Evolutionary Methods for Design, Optimization and Control*. J. Wiley & Sons, Chichester
14. Emmanouilidis C, Hunter A, MacIntyre J (2000) A multiobjective evolutionary setting for feature selection and a commonality-based crossover operator. *Proc. Congress on Evolutionary Computation (CEC'2000)*, 823–830. San Diego
15. Emmanouilidis C, Hunter A, MacIntyre J, Cox C (2001) A multi-objective genetic algorithm approach to feature selection in neural and fuzzy modeling. *Evolutionary Optimization* 3(1):1–26

16. Gustafson D, Kessel W (1985) Fuzzy identification of systems and its application to modeling and control. *IEEE Transactions on Systems, Man and Cybernetics* 15(1):116–132
17. Hamdani T, Won J, Alimi A, Karray F (2007) Multi-objective feature selection with NSGA-II. *Proc. of ICANNGA 2007*, 240–247
18. Huang B, Buckley B, Kechadi T (2010) Multi-objective feature selection by using NSGA-II for customer churn prediction in telecommunications. *Expert Systems with Applications* 37(5):3638–3646
19. Lac H, Stacey D (2005) Feature subset selection via multi-objective genetic algorithm. *Proc. IEEE Int. Joint Conf. on Neural Networks*, 1349–1354. IEEE Press, Piscataway
20. Michalewicz Z (1999) Genetic Algorithms + Data Structures = Evolution Programs, 3rd edition. Springer, Berlin
21. Miettinen K (1999) Nonlinear Multi-objective Optimization. Kluwer, New York
22. Nieto J, Alba E, Jourdan L, Talbi E (2009) Sensitivity and specificity based multi-objective approach for feature selection: Application to cancer diagnosis. *Information Processing Letters* 109:887–896
23. Oliveira L, Sabourin R, Bortolozzi F, Suen C (2002) Feature selection using multi-objective genetic algorithms for handwritten digit recognition. *Proc. 16th Int. Conf. on Pattern Recognation (ICPR' 02)*, 568–571. IEEE Press, Piscataway
24. Saha S, Ekbal A, Uryupina O, Poesio M (2011) Single and multi-objective optimization for feature selection in anaphora resolution. *Proc. 5th Int. Joint Conf. on Natural Language Processing*, 93–101
25. Sousa JMC, Kaymak U (2002) *Fuzzy Decision Making in Modeling and Control*. World Scientific and Imperial College, Singapore and UK
26. Srinivas N, Deb K (1994) Multiobjective optimization using nondominated sorting in genetic algorithms. *Evolutionary Computation* 2(3):221–248
27. Takagi T, Sugeno M (1985) Fuzzy identification of systems and its applications to modelling and control. *IEEE Trans. on Systems, Man and Cybernetics* 15(1):116–132
28. Unler A, Murat A (2010) A discrete particle swarm optimization method for feature selection in binary classification problems. *European Journal of Operational Research* 206:528–539
29. Vieira SM, Sousa JMC, Runkler TA (2010) Two cooperative ant colonies for feature selection using fuzzy models. *Expert Systems with Applications* 37(4):2714–2723
30. Wang C, Huang Y (2009) Evolutionary-based feature selection approaches with new criteria for data mining: A case study of credit approval data. *Expert Systems with Applications* 36:5900–5908
31. Wang X, Bandyopadhyay S, Xuan Z, Zhao X, Zhang M, Zhang X (2007) Prediction of transcription start sites based on feature selection using AMOSA. *Computational Systems Bioinformatics Conference{* 6:183–193
32. Waqas K, Baig R, Ali S (2009) Feature subset selection using multi-objective genetic algorithms. *Proc. 13th IEEE Multitopic Conference*, 1–6.