Multi-objective Genetic Algorithm Evaluation in Feature Selection

Newton Spolaôr[1,2], Ana Carolina Lorena[1], and Huei Diana Lee[2]

[1] Grupo Interdisciplinar em Mineração de Dados e Aplicações/Universidade Federal do ABC
Santo André, Brasil
[2] Laboratório de Bioinformática/Universidade Estadual do Oeste do Paraná
Foz do Iguaçu, Brasil
{newtonspolaor,aclorena,hueidianalee}@gmail.com

Abstract. Feature Selection may be viewed as a search for optimal feature subsets considering one or more importance criteria. This search may be performed with Multi-objective Genetic Algorithms. In this work, we present an application of these algorithms for combining different filter approach criteria, which rely on general characteristics of the data, as feature-class correlation, to perform the search for subsets of features. We conducted experiments on public data sets and the results show the potential of this proposal when compared to mono-objective genetic algorithms and two popular filter algorithms.

Keywords: filter feature selection, feature importance measures, multi-objective genetic algorithms

1 Introduction

Enormous volume of data has been collected, due to the development of technology, and organized in Databases (DB). Computational processes like Data Mining (DM) may be applied in order to analyze these DB. DM enables the construction of logical hypothesis (models) from data, potentially extracting useful knowledge for specialists, that can be used as a second opinion in decision making processes [18].

The DM process is mainly composed of pre-processing, pattern extraction and pos-processing. Pre-processing involves the proper representation of the data into forms like attribute-value, in which lines and columns represent, respectively, examples and features (attributes, characteristics) of the data set. Other pre-processing tasks include cleaning the data and Feature Selection (FS), which is the focus of this work. The pattern extraction phase involves the construction of models from data, using, for example, Machine Learning algorithms. The obtained models are evaluated and consolidated by the specialists at the end of the process (pos-processing).

FS may be formulated as a search for an optimal subset of features in a DB, in which each state of the search space represents a possible subset of features [25]. The optimality of this subset may be estimated according to a maximization or minimization function of one or more measures (criteria) of importance of features. Applying FS allows mapping the original data to a projection in which the examples are described by part of the features. This leads to a dimensional reduction of the data set. Models constructed using these projections may have lower complexity and potentially superior or equivalent quality when compared to those generated using the original data. In addition, FS may help on a better comprehension of the domain, by maintaining only the features with a good ability, according to some importance criterion, to describe the inherent patterns within the data and helps to reduce the effects of the curse of dimensionality [25].

Searching related to FS is usually a combinatorial process [7], precluding the investigation of all subsets. This is one of the motivations to apply heuristic search methods such as Genetic Algorithms (GA) [27] in this process. Furthermore, it may be interesting to find subsets of features that optimize different importance criteria, leading to the motivation of using Multi-objective Optimization strategies (MO) [6]. In the literature, there are many applications of Multi-objective Genetic Algorithms (MOGA) in different areas and tasks, including FS [36, 5, 13, 44, 37, 3, 19, 42].

The objective of this work is to evaluate the application of MOGA to FS based on different filter importance criteria. The performed experiments investigate distinct combinations of these criteria, what is not performed in works related to the filter approach [5, 37, 3, 42]. This work also differentiates from previous work [35, 36, 5, 29, 33, 42, 34] by including a comparative evaluation of the MOGA with distinct mono-objective GA, where each GA optimizes one filter importance criterion individually, and two popular filter FS algorithms. The selected subsets are evaluated through the construction of models using two pattern extraction algorithms in nine benchmark data sets. The predictive ability of these models is statistically compared to the performance of models built using all features.

This study is part of the Intelligent Data Analysis project (IDA) [35, 36, 23], which is developed in a partnership among the "Universidade Federal do ABC" (UFABC), the "Laboratório de Bioinformática/Universidade Estadual do Oeste do Paraná" (LABI/UNIOESTE), the "Laboratório de Inteligência Computacional/Universidade de São Paulo" (LABIC/USP) and the "Serviço de Coloproctologia/Universidade Estadual de Campinas" (UNICAMP).

This work is organized as follows: in Section 2 concepts related to FS and also importance measures used in the MOGA are described. The complete proposal is described in Section 3 and its evaluation, using nine data sets from a public repository, is presented in Section 4. Final considerations are made in Section 5.

2 Feature Selection

Feature Selection may be viewed as a dimensional reduction process of a data set in order to maintain only its most important features according to some criterion. The importance criteria are usually based on the principles of relevance and non-redundancy among features [15, 23, 17] and may be organized into measures of consistency, dependency, distance, information and precision [25]. In this work, one measure of each of the mentioned categories was used with the MOGA, excluding precision. These measures were chosen from work related to filter Feature Selection [1, 42, 20, 38, 28].

All considered data sets are for supervised learning and related to classification problems. In these data sets, each example (or case) has an associated label (or class) and the objective is to construct predictive models, which are capable of predicting the label of new cases previously unknown. The data set is composed of n pairs $(\mathbf{x}_i, y_i)$, in which $\mathbf{x}_i = (x_i(1), \ldots, x_i(m))$ represents an example with m features and y_i corresponds to its class. The exclusion of a relevant feature F_1 results in a worse predictive performance of the correspondent classification model. Two features F_2 and F_3 are said to be non-redundant when they are not significantly correlated.

It is relevant to mention that the feature importance estimation may be considered in two ways: individually or in subsets. Nevertheless, the individual evaluation methods are incapable of removing redundant features, as they may present the same relevance [17]. For this reason, in this work we considered FS in subsets. Individual importance evaluations are combined into a unique resultant value, representing the subset of features as a whole.

The importance of features may be viewed according to the interaction with the pattern extraction algorithm [25]. In the wrapper approach, a pattern extraction algorithm, which will be used later for the construction of models, is considered for selecting features. For each subset, a model using this specific algorithm is constructed and evaluated. In the filter approach, feature subsets are evaluated before the pattern extraction step, and considers general characteristics of the data, such as statistical measures, to select the important features. A third approach is the embedded, in which the process of selecting features is performed internally by the pattern extraction algorithm, as in the case of decision trees [31].

Importance Measures Importance measures inspired in the concept of consistency value, for example, chooses subsets of features that minimize the occurrence of inconsistent pairs of examples in discretized data sets, that is, which present identical values in each feature but different labels [1]. The Inconsistent Example Pairs (IP) measure identifies the inconsistency rate by the ration of the number of inconsistent pairs of examples and the total number of pairs of examples.

Correlation measures enable a redundancy analysis of the data set when estimating the prediction capability of a feature. The Attribute Class Correlation

(AC) [38] exemplifies this category, and is described by Equation 1, where w_i will be 1 if i is selected and 0 otherwise; $\phi(.,.) = 1$ if j_1 and j_2 have distinct labels or -0.05 otherwise. $|.|$ denotes the module function. The formulation of $C(i)$ demonstrates that this measure highlights feature values that show the most distinct values for examples of different classes.

$$AC = \left(\sum w_i C(i)\right) / \left(\sum w_i\right) \tag{1}$$
$$\text{where } C(i) = \frac{\sum_{j_1 \neq j_2} |x_{j1}(i) - x_{j2}(i)| \, \phi(\mathbf{x}_{j1}, \mathbf{x}_{j2})}{n(n-1)/2}.$$

The Inter-Class Distance measure (IE) [42] estimates the existent separability between classes when the set of examples is described only by the investigated subset of features. The separability maximization may be useful to generate classification models, as the differentiation of diverse patterns is favored. Equation 2 presents IE, where $\mathbf{p}$ is the central example (centroid) of a data set with k classes, $d(.,.)$ denotes the Euclidean distance, and $\mathbf{p}_r$ and n_r represent, respectively, the central example and the number of examples in class r.

$$IE = \frac{1}{n} \sum_{r=1}^{k} n_r d(\mathbf{p}_r, \mathbf{p}). \tag{2}$$

The Laplacian Score (LS) [20] is also based on distance and is inspired by the possibility of identifying examples with affinity when they are relatively next to each other. In classification, for example, this behavior is potentially observed among instances of the same label, highlighting the importance of modeling the related local structure. Herewith, LS proposes building a nearest neighbor graph, in which each node corresponds to a distinct example and the nearest examples are connected by arcs. The S weight matrix of this graph is considered in Equation 3, with $\mathbf{x}(i) = [x_1(i), x_2(i), \cdots, x_n(i)]^T$ and $\mathbf{1} = [1, \cdots, 1]^T$. This formula includes the matrices $D = diag(S\mathbf{1})$, in which $diag(.)$ extracts the diagonal matrix, and the Laplacian Graph [8] $L = D - S$.

$$LS(i) = \frac{\tilde{\mathbf{x}}(i)^T L \tilde{\mathbf{x}}(i)}{\tilde{\mathbf{x}}(i)^T D \tilde{\mathbf{x}}(i)} \tag{3}$$
$$\text{where } \tilde{\mathbf{x}}(i) = \mathbf{x}(i) - \frac{\mathbf{x}(i)^T D\mathbf{1}}{\mathbf{1}^T D\mathbf{1}}\mathbf{1}.$$

Information based measures may be applied to reduce the uncertainty associated to the investigated problem. Representation Entropy (RE) [28], for example, enables the investigation of the information distribution among features and, consequently, to estimate the involved redundancy [41]. RE is presented by Equation 4, in which the λ_i eigenvalues are extracted from a covariance matrix of features of m order.

$$RE = -\sum \tilde{\lambda}_i \log \tilde{\lambda}_i \quad (4)$$
$$\text{where } \tilde{\lambda}_i = \frac{\lambda_i}{\sum \lambda_i}.$$

Precision measures consider information like the accuracy rate of the model in the classification of examples described by a subset of features or other estimate of the models' quality. Usually these measures are related to the wrapper approach and are not considered in this work.

3 MOGA in Feature Selection

MOGA offers the combination of GA and MO for the solution of search and optimization problems with multiple objectives [9]. Searching for subsets of important features in a data set can be considered a multi-objective task, since there are multiple criteria for measuring their importance, and each one of them considers different aspects of data [1, 42, 20, 38, 28].

With the goal of ranking risk factors related to premature births, in [42] the Non-dominated Sorting Genetic Algorithm (NSGA-II) [12] MOGA was used in FS relating, through the Pareto strategy, importance measures IE, AC and Intra-Class Distance. Some results are superior to those of models built using all features and also of other techniques for FS.

In [5] the NSGA-II algorithm is applied in supervised and semi-supervised FS in data sets of hyperspectral images. Two measures were optimized simultaneously: discrimination between classes and spatial invariance of the features. In general, the results obtained demonstrate a superior performance of MOGA over mono-objective GA.

The same MOGA is used in [3] for FS in the classification of microarray gene expression profiles. Because of the nature of these data, which generally have few examples and many features, the classification task becomes more complex, motivating FS. The importance measures of cardinality and ability to discriminate examples were jointly optimized. Experimentally, there were accuracy gains when compared to a mono-objective GA and other techniques for FS.

The inter and intra correlation measures proposed by [37] were applied for FS in data sets for the analysis of credit risk, using the algorithm NSGA-II. Experimentally, the model built using the features selected by the MOGA had a better performance than those models generated using all features and also using features selected by mono-objective GA and the Relief technique [40].

This work differs from previous work by investigating some combinations of filter importance measures belonging to different categories. The individuals were encoded using a binary chromosome with m genes, each of which corresponds to a distinct feature. A gene i with value 1 represents the selection of its respective feature, while the value 0 indicates its exclusion. A randomly initialized population of individuals is then evolved until a number of generations is reached. The NSGA-II MOGA was used, as in the previous related work.

The importance measures used as objective functions to be optimized are those discussed in Section 2, which belong to the classes: consistency, dependence, distance and information. The aim is to exploit complementarities between representatives of measures from different categories. We investigated the optimization of these measures in pairs always involving IE and some other measure. This choice is based on previous results presented in [35, 36], where the combinations involving the IE measure were more successful experimentally.

Experiments with three objectives led to a greater computational cost and little gains in other aspects, such as in the reduction obtained on the subsets of selected features. Furthermore, it is known that MOGA based on the Pareto theory do not scale well for optimization problems with more than three objectives [21]. For these reasons, only pairs of importance measures are considered in this work.

LS is the only measure used in the study which evaluates the importance of each feature individually. In its case, we used the average value calculated for each selected feature (with value 1 on a chromosome s). The other measures are calculated for each chromosome, using the subset of features represented by genes with value 1.

We used the one-point crossover, bit-flip mutation and binary tournament [27] in the MOGA. NSGA-II returns a set of optimal solutions, representing different tradeoffs between the objectives considered. We used the Compromise Programming (CP) technique [43] to select a single solution from this set due to its relative simplicity.

4 Experimental Evaluation

We applied the NSGA-II for FS described in the previous section in nine data sets from the UCI repository[3] [2]: Australian (A), Dermatology (D), Ionosphere (I), Lung cancer (L), Sonar (S), Soybean small (Y), Vehicle (V), Wine (W) and Wisconsin breast cancer (B). All features in these data sets are numerical and have continuous or discrete values. Table 1 presents, for each data set, the Majority Class Error (MCE) rate, which corresponds to the error rate obtained by classifying all data in the majority class, and the number (♯) of examples, features and classes.

Table 1. Data sets information

	A	**D**	**I**	**L**	**S**	**Y**	**V**	**B**	**W**
♯Examples	690	358	351	32	208	47	846	569	178
♯Features	14	34	34	56	60	35	18	30	13
♯Classes	2	6	2	3	2	4	4	2	3
MCE	44.49	68.99	35.9	59.37	46.63	63.83	74.23	37.26	60.11

[3] Supported by the Turing Institute in Glasgow (Vehicle)

We used the NSGA-II implementation available in the Platform and programming language Independent interface for Search Algorithms (PISA) [4], with the following parameters: $\alpha = 50$, $\mu = 50$, $\lambda = 50$, $crossover\ rate = 0.8$, $mutation rate = 0.01$, $stopping criterion = 50 generations$. The parameters α, μ and λ correspond, respectively, to the population size and the number of parents and children individuals after reproduction. Their values were defined based on related work. Another tool used in the implementations was the GNU Scientific Library (GSL)[4], which enables the implementation of the covariance matrices associated with the RE measure.

As previously mentioned, we have investigated multi-objective combinations involving the IE measure and each of the other four importance measures described in Section 2, resulting in four distinct multi-objective settings. The evaluation of the subsets of features selected by the MOGA in each multi-objective setting was performed by building classification models using projections of the data sets containing the selected features. Classification algorithms J48, an implementation of C4.5 [31], and Support Vector Machines (SVM) [10], from the Weka tool [40], were used in the induction of classifiers. Their parameter values were kept default. These classifiers were selected due to: the relatively low number of parameters, in the case of J48; and robustness to high dimensional data, in the case of SVM.

We also implemented five mono-objective GA for FS, each of them using one of the importance measures discussed in this work as fitness function. The same binary encoding and genetic operators from MOGA were used. The results of the FS algorithms Correlation-based Feature Subset Selection (CFS) [16] and Consistency Subset Eval (CSE) [24] from literature are also presented. CFS chooses subsets of features highly correlated with the class and that have low inter-correlation, while CSE analyzes the consistency of the data projections obtained using different subsets of features.

CFS and CSE filter algorithms come from the Weka tool and were employed with default parameter values. The mono-objective GA's parameters number of generations, population size and probabilities of crossover and mutation were changed in order to be identical to those used for NSGA-II. The population size was set to 50 and the seed of GA, as for NSGA-II, was set randomly for each run. The classification models defined using all the features in each data set were included as baselines (c_a). In LS we used as neighborhood of each example its five nearest neighbors in terms of distance.

In the experiments, each data set d was initially divided according to Stratified Cross-Validation (SCV) into 10 folds, which leads to 10 pairs of training and test sets. Because of MOGA stochasticity, it was executed five times for each training partition f_i, and for each multi-objective setting m_s. One unique subset of features is identified for each MOGA run, using CP. This results in five subsets of features per multi-objective setting. This enables the generation of five different projections of the data partition f_i. After training classification models using the five projections and evaluating them on their corresponding

[4] http://www.gnu.org/software/gsl

test partitions, 50 accuracy rates are obtained. Similarly, we counted up the Percent of Reduction (PR) in the amount of original features for each run. The mean values of these evaluation measures are reported for each setting m_s. The mono-objective GA are subjected to a similar procedure, while the other FS algorithms and the baselines are evaluated by the mean values obtained in a unique run for each of the 10 folds of d.

4.1 Results

For each data set, we show in Figures 1 and 2 the PR and the accuracy rate of the J48 and SVM models for each FS algorithm when related to those rates obtained when using all features (c_a). Therefore, if a classifier c_i and c_a accuracy rates are, respectively, 85.36% and 83.48%, the graph displays for c_i the result of the ratio between these rates (1.02). The horizontal line corresponds to the point where this ratio reaches the value 1 in each graph, that is, when the rates are equal for both c_i and c_a. The dark bars represent PR and the light bars show the accuracy rates of the classifiers built using the features selected by each FS algorithm. Therefore, if the top of a light bar is above the baseline line, the accuracy of the model represented is higher than that of the baseline. The black bars will never exceed the baseline line, because PR is always less than 100% of the original number of features per data set. An aggressive PR is identified when the top of its bar is close to the baseline line.

We noticed that, in general, the magnitude of the standard deviations of the accuracy rates for each c_i had no strong discrepancy to those achieved by c_a. Importantly, the FS embedded in J48 was not investigated in this work, therefore all PR illustrated refer to subsets of features identified by the FS algorithms evaluated.

Since we do not have assurance of normality, we employed the non-parametric Kruskall-Wallis test [22] separately for each data set to identify statistical differences at 95% of significance level between each of the algorithms evaluated and the baseline in terms of accuracy rate. Using a unique baseline implies in less statistical comparisons, softening the multiplicity effect [32]. Models with higher and lower statistical accuracy performance when compared to c_a are highlighted in the graphs, respectively, with a triangle pointing up and down. Models that had accuracy not statistically lower than that of c_a, while providing PR greater than 50%, are highlighted with a circle. Table 2 summarizes these information, presenting the total number of models significantly better (in brackets) and with no significant difference when compared to c_a, for each classifier. It also shows the mean PR and standard deviation (in parentheses) by FS algorithm for all data sets.

4.2 Discussion

We identified 7 models (3.5% of the total) with significant superiority and 135 models (68.2%) with no significant difference of results when compared to the baseline, of which 46 had PR higher than 50%. A reduction in the number of

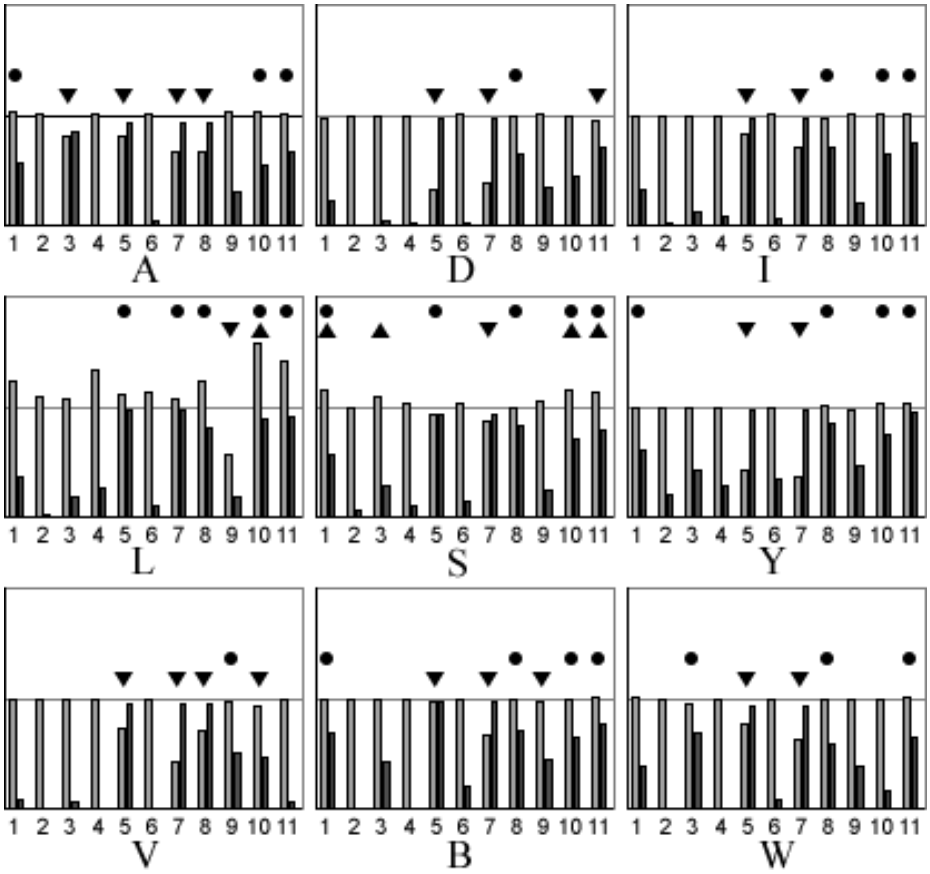

Fig. 1. J48 models generated after applying: (1) IE + AC, (2) IE + IP, (3) IE + LS, (4) IE + RE, (5) AC, (6) IE, (7) LS, (8) IP, (9) RE, (10) CFS and (11) CSE.

Table 2. Models not statistically inferior to c_a.

	IE + AC	IE + IP	IE + LS	IE + RE	AC	IE
J48	8 (1)	9 (0)	7 (1)	9 (0)	2 (0)	9 (0)
SVM	8 (0)	9 (0)	7 (0)	9 (0)	1 (0)	9 (0)
PR	42.07 (20)	3.18 (6.7)	33.84 (28.5)	8.53 (11.7)	95.17 (2.1)	10.94 (10.7)

	LS	IP	RE	CFS	CSE
J48	1 (0)	7 (0)	7 (0)	6 (2)	7 (1)
SVM	1 (0)	3 (0)	6 (0)	7 (1)	3 (1)
PR	95.24 (1.9)	78.22 (12.3)	33.82 (11.7)	58.11 (21)	69.41 (25.7)

features with the maintenance or improvement in accuracy when compared to the baseline is important, because it allows reducing the computational cost of

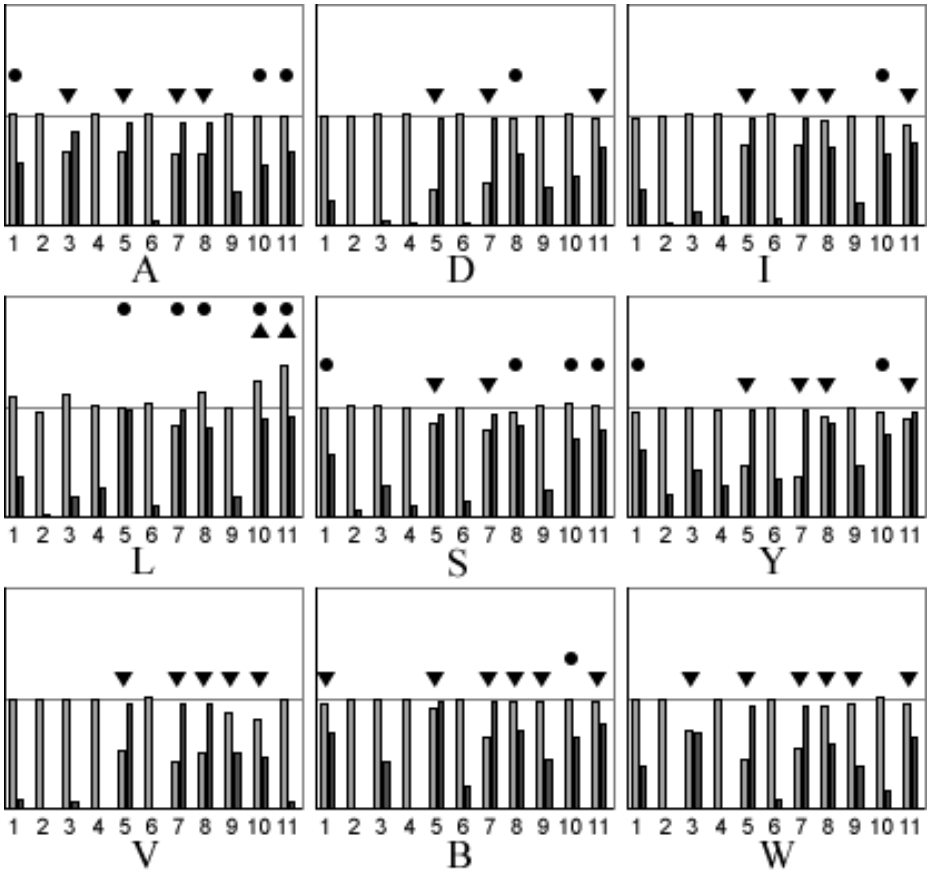

Fig. 2. SVM models gerarated after applying: (1) IE + AC, (2) IE + IP, (3) IE + LS, (4) IE + RE, (5) AC, (6) IE, (7) LS, (8) IP, (9) RE, (10) CFS and (11) CSE.

the classification model induced and can also contribute to improve its comprehensibility [25, 38]. However, most of these occurrences are concentrated in experiments related to the multi-objective setting IE + AC and to the CFS and CSE algorithms. This behavior is reinforced by the results shown in Table 2, where these algorithms showed PR greater than 40% and are associated to all cases of statistical superiority when compared to the baseline. We also indentified in several experiments, mainly those related to mono-objective GA using measures AC and LS, that a too aggressive dimensional reduction led to a predictive performance statistically inferior to that of c_a.

The MOGA based on the IE + AC measures stood out in comparison to the other settings by allowing the generation of models with lower computational complexity and predictive performance statistically similar to the baseline in different data sets. Additionally, in the Sonar data set it was possible to ob-

tain a J48 model with superior performance when compared to c_a. These results reinforce previous experimental evidence [35, 36] of the importance of selecting features that maximize the separability between classes for supervised classification problems. The combined use of IE with other measures of importance can contribute, for example, to select just one feature from two or more features that are equivalent in terms of separability.

The mix between IE and AC explores the positive aspects of measures belonging to distinct categories, which can be observed in the results of their individual optimization in the mono-objective GA. The isolated use of the IE measure results in many models with high predictive performance, but that maintain most of the features in the data sets. In many cases there is no dimensional reduction at all with the isolated use of this measure. On the other hand, the GA that uses the AC fitness function made aggressive reductions in the number of features and generated models that also stand out for their predictive power, although there are cases of significant losses in terms of accuracy.

IE + LS setting explores in a smaller scale the aggressiveness of the LS measure in reducing the percentage of features, which is similar to that obtained by the AC criterion. This behavior is emphasized, in general, by observing that this combination presents a greater number of models with accuracies statistically lower than those of the baseline models when compared to other multi-objective settings. A possible justification for this fact is that both measures belong to the distance category, what is not observed in the other combinations investigated. Thereafter, their combination do not enjoy the benefits of MO for FS regarding different categories of importance.

In addition, the LS measure is dependent on a parameter for the construction of the weight matrix of the graph that models the local structure of the data. This parameter was set in the experiments with a unique value, motivating further studies for the investigation of other values. This study may also contribute to prevent the occurrence of cases in which no feature is selected by the mono-objective GA with the LS measure, which was a specific behavior of this criterion in previous experiments which, for instance, show the problem of division by zero in Equation 3.

It is interesting to notice that the combinations between the IE measure and measures IP and RE tends to maintain all features in several cases. In fact, it was found experimentally that the measures RE and IP in general are relatively more conservative than the criteria AC and LS regarding PR, which may have contributed to the fact that these measures reached satisfactory accuracy rates in some mono-objective experiments. For monotonic criteria as IP, conservatism can be explained because a larger number of selected features may allow to define more logical hypothesis [1]. In general, it appears that the joint optimization of conservative measures, such as IE, IP and RE tends to generate models that keep that characteristic. Specifically in MOGA IE + IP and IE + RE, the strong conservatism of the IE measure prevailed over the mild conservatism of IP and RE measures.

An analysis of the results obtained by CFS and CSE shows that these algorithms are competitive with the GA investigated. The MOGA optimizing IE + AC is that which most closely approximates the results of these algorithms considering predictive performance. The number of models that are highlighted with a circle or statistically better than c_a after the application of these techniques is higher than that observed for all mono-objective and multi-objective GA, while the average PR for each one of them in the nine data sets is higher than 50%. The CSE algorithm specifically exhibits the disadvantage of presenting a larger number of models with results statistically lower than the baseline when compared to IE + AC MOGA and CFS. Importantly, both FS algorithms, like all MOGA, present in most experiments a larger number of models with no statistical difference when compared to the baseline.

We also noticed that the use of J48 provided the occurrence of a larger total number of models with statistical superiority to c_a than the SVM. It appears that the accuracy rate of models generated from projections of features is influenced by the classification technique used afterwards. This influence may have led to the fact that a FS algorithm underperforming c_a for a particular classification technique can be superior for other technique. Future work shall investigate the predictive behavior of other classification techniques with and without FS. Some initial experiments have been performed and have confirmed those observations.

In future we also plan to combine in a MOGA the IE and the importance measure of CFS (original or altered as in [26, 30]). Herewith, it would be possible to perform FS considering both distance and dependency, as in IE + AC, using measures that have been investigated in recent studies [11, 14, 35, 36, 42]. In fact, the AC dependence measure explores only relevance, selecting features most correlated with the data labels. Since the measure used in CFS also considers the correlation between features, it also addresses the non-redundancy aspect.

Therefore, the method used in this work supports the implementation of different measures of importance of features, including those from algorithms CFS and CSE. This flexibility already enabled the investigation of labeled data with numerical feature values using different combinations of six criteria, taken in pairs or triplets [35, 36]. Besides, some importance measures are also flexible. Measures such as IE can be used for FS in data sets with categorical features by using other distance metrics [39], while the LS and RE criteria are applicable to unlabeled data. Another advantage of the MOGA is its ability to return multiple solutions (various subsets of features). Only one of them was selected with the CP technique in our work, but others could be selected using different techniques or they could be even combined.

All FS algorithms investigated in this work belong to the filter approach. It was possible to build many models with similar accuracy to those of classifiers that use all features, while using lower numbers of features. For J48, for example, this could lead to the obtainment of decision trees with fewer nodes and can result in more understandable decision rules. These advantages are achieved with a computational cost potentially lower than would be obtained with algorithms that employ wrapper measures [25].

5 Conclusion

This work presented an evaluation of a MOGA for FS in labeled data sets, including different multi-objective configurations of features' importance measures. Their results were also compared to those of GA that optimize each importance measure individually and two popular FS techniques available in a free tool. In the experimental results and discussions we observed a prominence of the combination of IE and AC measures, coming from categories based on distance and dependency, respectively, and of the two filter algorithms from literature, which allowed to obtain different models with reduced number of features and good predictive performance.

The good results of the IE + AC MOGA in this study suggest that combining measures belonging to different categories of importance is interesting for FS in labeled data. The multi-objective optimization of these measures enables the identification of relevant features both in terms of separability between examples and correlation between features and the class. It should be also worth to analyse the degree of complementarity of these measures.

As future work we aim to combine the measures IE and that of the CFS algorithm in NSGA-II and also compare the results of different classifiers when using the subsets of features. It is also interesting to investigate the application of measures such as IE and IP in data sets with categorical features and perform experiments with the LS and RE criteria in unlabeled data sets. We can still observe the influence of different parameter values for the LS importance measure and to compare the investigated MOGA to other multi-objective metaheuristics, as well as to wrapper FS approach.

Acknowledgements To UFABC, CAPES, FAPESP and CNPq for the support received for this work and to the staff of the IDA project for the cooperation.

References

1. Arauzo-Azofra, A., Benitez, J.M., Castro, J.L.: Consistency measures for feature selection. Journal of Intelligent Information Systems 30(3), 273–292 (2008)
2. Asuncion, A., Newman, D.: UCI machine learning repository. http://www.ics.uci.edu/∼mlearn/MLRepository.html (2007)
3. Banerjee, M., Mitra, S., Banka, H.: Evolutionary rough feature selection in gene expression data. IEEE Transactions on Systems Man and Cybernetics 37(4), 622–632 (2007)
4. Bleuler, S., Laumanns, M., Thiele, L., Zitzler, E.: PISA — a platform and programming language independent interface for search algorithms. In: Evolutionary Multi-Criterion Optimization. pp. 494–508 (2003)
5. Bruzzone, L., Persello, C.: A novel approach to the selection of spatially invariant features for the classification of hyperspectral images with improved generalization capability. IEEE transactions on geoscience and remote sensing 47, 3180–3191 (2009)

6. Bui, L.T., Alam, S.: An Introduction to Multiobjetive Optimization. Information Science Reference (2008)
7. Charikar, M., Guruswami, V., Kumar, R., Rajagopalan, S., Sahai, A.: Combinatorial feature selection problems. In: Annual Symposium on Foundations of Computer Science. pp. 631–640 (2000)
8. Chung, F.: Spectral Graph Theory. AMS (1997)
9. Coello, C.A.C.: Evolutionary multi-objective optimization: a historical view of the field. Computational Intelligence Magazine pp. 28–36 (2006)
10. Cristianini, N., Shawe-Taylor, J.: Support Vector Machines and other Kernel-Based Learning Methods. Cambridge University Press (2000)
11. Danger, R., Segura-Bedmar, I., Martínez, P., Rosso, P.: A comparison of machine learning techniques for detection of drug target articles. Journal of Biomedical Informatics pp. 1–12 (2010)
12. Deb, K., Agrawal, S., Pratap, A., Meyarivan, T.: A fast elitist non-dominated sorting genetic algorithm for multi-objective optimization: Nsga-ii. In: Schoenauer, M., Deb, K., Rudolph, G., Yao, X., Lutton, E., Merelo, J., Schwefel, H.P. (eds.) Parallel Problem Solving from Nature, pp. 849–858. Springer Berlin (2000)
13. Dessì, N., Pes, B.: An evolutionary method for combining different feature selection criteria in microarray data classification. Journal of Artificial Evolution and Applications pp. 1–10 (2009)
14. Duangsoithong, R., Windeatt, T.: Correlation-based and causal feature selection analysis for ensemble classifiers. In: Artificial Neural Networks in Pattern Recognition. pp. 25–36 (2010)
15. Dy, J.G.: Unsupervised feature selection. In: Liu, H., Motoda, H. (eds.) Computational Methods of Feature Selection, pp. 19–39. Chapman & Hall/CRC (2008)
16. Hall, M.A.: Correlation-based Feature Selection for Machine Learning. Phd thesis, University of Waikato (1999)
17. Hall, M.A.: Correlation-based feature selection for discrete and numeric class machine learning. In: International Conference on Machine Learning. pp. 359–366 (2000)
18. Han, J., Kamber, M.: Data mining: concepts and techniques. Morgan Kaufmann (2006)
19. Handl, J., Kell, D.B., Knowles, J.: Multiobjective optimization in bioinformatics and computational biology. IEEE/ACM Transactions on Computational Biology and Bioinformatics pp. 279–292 (2007)
20. He, X., Cai, D., Niyogi, P.: Laplacian score for feature selection. In: Advances in Neural Information Processing Systems. pp. 507–514 (2005)
21. Jaimes, A.L., Coello, C.A., Barrientos, J.E.U.: Online objective reduction to deal with many-objective problems. In: International Conference on Evolutionary Multi-Criterion Optimization. pp. 423–437 (2009)
22. Kruskal, W., Wallis, W.A.: Use of ranks in one-criterion variance analysis. American Statistical Association 47, 583–621 (1952)
23. Lee, H.D., Monard, M.C., Wu, F.C.: A fractal dimension based filter algorithm to select features for supervised learning. In: Advances in Artificial Intelligence. pp. 278–288 (2006)
24. Liu, H., Setiono, R.: A probabilistic approach to feature selection - a filter solution. In: International Conference on Machine Learning. pp. 319–327 (1996)
25. Liu, H., Motoda, H.: Computational Methods of Feature Selection. Chapman & Hall/CRC (2008)
26. Lutu, P.E.N., Engelbrecht, A.P.: A decision rule-based method for feature selection in predictive data mining. Expert Systems with Applications 37(1), 602–609 (2010)

27. Mitchell, M.: An introduction to genetic algorithms. MIT Press (1998)
28. Mitra, P., Murthy, C.A., Pal, S.K.: Unsupervised feature selection using feature similarity. IEEE Transactions on Pattern Analysis and Machine Intelligence 24(3), 301–312 (2002)
29. Neshatian, K., Zhang, M.: Pareto front feature selection: using genetic programming to explore feature space. In: Proceedings of the 11th Annual conference on Genetic and evolutionary computation. pp. 1027–1034 (2009)
30. Nguyen, H., Franke, K., Petrovic, S.: Improving effectiveness of intrusion detection by correlation feature selection. In: International Conference on Availability, Reliability and Security. pp. 17–24 (2010)
31. QuinLan, J.: C4.5: Programs for Machine Learning. Morgan Kaufmann (1993)
32. Salzberg, S.L.: On comparing classifiers: Pitfalls to avoid and a recommended approach. Data Mining and Knowledge Discovery 1, 317–328 (1997)
33. Santana, L.E.A., Silva, L., Canuto, A.M.P.: Feature selection in heterogeneous structure of ensembles: a genetic algorithm approach. In: International Joint Conference on Neural Networks. pp. 1491–1498 (2009)
34. Shon, T., Kovah, X., Moon, J.: Applying genetic algorithm for classifying anomalous tcp/ip packets. Neurocomputing 69, 2429–2433 (2006)
35. Spolaôr, N., Lorena, A.C., Lee, H.D.: Seleção de atributos por meio de algoritmos genéticos multiobjetivo (in portuguese). In: Workshop on MSc Dissertation and PhD Thesis in Artificial Intelligence. pp. 1–10 (2010)
36. Spolaôr, N., Lorena, A.C., Lee, H.D.: Use of multiobjective genetic algorithms in feature selection. In: IEEE Brazilian Symposium on Artificial Neural Network. pp. 1–6 (2010)
37. Wang, C.M., Huang, Y.F.: Evolutionary-based feature selection approaches with new criteria for data mining: A case study of credit approval data. Expert Systems with Applications 36(3), 5900–5908 (2009)
38. Wang, L., Fu, X.: Data Mining With Computational Intelligence. Springer (2005)
39. Wilson, D.R., Martinez, T.R.: Improved heterogeneous distance functions. Journal of Artificial Intelligence Research 6, 1–34 (1997)
40. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann (2005)
41. Yan, W.: Fusion in multi-criterion feature ranking. In: International Conference on Information Fusion. pp. 01–06 (2007)
42. Zaharie, D., Holban, S., Lungeanu, D., Navolan, D.: A computational intelligence approach for ranking risk factors in preterm birth. In: International Symposium on Applied Computational Intelligence and Informatics. pp. 135–140 (2007)
43. Zeleny, M.: An introduction to multiobjetive optimization. In: Cochrane, J.L., Zeleny, M. (eds.) Multiple criteria decision making, pp. 262–301. University of South Carolina Press (1973)
44. Zhu, Z., Ong, Y.S., Kuo, J.L.: Feature selection using single/multi-objective memetic frameworks. In: Goh, C.K., Ong, Y.S., Tan, K.C. (eds.) Multi-Objective Memetic Algorithms, pp. 111–131. Springer-Verlag (2009)