Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 102 (2016) 51 – 58

Procedia Computer Science

12th International Conference on Application of Fuzzy Systems and Soft Computing, ICAFS 2016, 29-30 August 2016, Vienna, Austria

Fuzzy criteria in multi-objective feature selection for unsupervised learning

Fuyu Cai[a], Hao Wang[b], Xiaoqin Tang[a], Michael Emmerich[b], Fons J. Verbeek[a,*]

[a]*Section Imaging and Bioinformatics, LIACS, Leiden University, Niels Bohrweg 1, 2333 CA, Leiden, The Netherlands*
[b]*Section Algorithms and Software Technology, LIACS, Leiden University, Niels Bohrweg 1, 2333 CA, Leiden, The Netherlands*

Abstract

Feature selection in which most informative variables are selected for model generation is an important step in pattern recognition. Here, one often tries to optimize multiple criteria such as discriminating power of the descriptor, performance of model and cardinality of a subset. In this paper we propose a fuzzy criterion in multi-objective unsupervised feature selection by applying the hybridized filter-wrapper approach (FC-MOFS). These formulations allow for an efficient way to pick features from a pool and to avoid misunderstanding of overlapping features via crisp clustered learning in a conventional multi-objective optimization procedure. Moreover, the optimization problem is solved by using non-dominated sorting genetic algorithm, type two (NSGA-II). The performance of the proposed approach is then examined on six benchmark datasets from multiple disciplines and different numbers of features. Systematic comparisons of the proposed method and representative non-fuzzified approaches are illustrated in this work. The experimental studies show a superior performance of the proposed approach in terms of accuracy and feasibility.

Peer-review under responsibility of the Organizing Committee of ICAFS 2016

Keywords: feature selection; fuzzy criteria; unsupervised learning; multi-objective optimization

* Corresponding author. Tel.: +31-(0)624236531. fax: +31 - (0)71 527 6985
E-mail address: f.j.verbeek@liacs.leidenuniv.nl

Peer-review under responsibility of the Organizing Committee of ICAFS 2016
doi:10.1016/j.procs.2016.09.369

1. Introduction

Feature selection (FS), in some areas also referred to as dimensionality reduction, deals with selection of one or several optimal sets of attributes that are necessary and/or essential for the recognition process. The challenge of FS is to decide a minimum subset of features with little or no loss of classification/clustering accuracy. This can be formulated as a multi-objective optimization (MOO) problem. The task is the selection of relevant features, elimination of redundant features, and minimization of selected set cardinality. To date, a range of MOO-based FS techniques have been reported[14]. Cross-applications the related FS approaches can be categorized into four groups:

- Filter-supervised, i.e. class-labels known: features are selected based on their discriminating power with respect to the target classes.
- Wrapper-supervised, i.e. class labels known: subsets of features are evaluated from classification, at the point where comparison of resulting labels and actual labels occurs.
- Filter-unsupervised, i.e. class-labels unknown: features are ranked from the performance histogram of all feature dimension vectors and one or several criteria are chosen for deciding a group of features.
- Wrapper-unsupervised, i.e. class-labels unknown: computation of the subset of features is applied in terms of the performance of a clustering algorithm. In this case, tuning of parameters in clustering process will contribute in obtaining an acceptable subset of features.

The search for proper supervised predictors can usually be regarded as a pursuit for optimization, where the number of wrong-predicted operators for a known dataset should be minimized[1]. However, figuring out a similar criterion for validation in unsupervised schemas is a difficult task[2]. It cannot be relied upon that a new-found pattern obtained by optimizations resulting from an unsupervised algorithm, is able to decide if a given pattern is trustful or not. To some extent, the validity of pattern discovery is depended on a priori knowledge and intentions of decision makers. This brings us to the assumption that one often desires to employ unsupervised learning schemas in order to produce several candidate solutions for users. Additionally, some tasks in FS, cover inherent data groups and thereby omit features which might reveal the nature of hidden patterns. Therefore, the unsupervised-based multi-objective heuristic optimization algorithm is becoming an attractive approach, that has been given and increasing attention this decade.

There has been reported on development of evolutionary algorithms for multi-objective (MOEA) for unsupervised feature selection[3]. Oliveira, et al.[4] proposed a Pareto-based approach to generate a so-called Pareto-optimal front in a supervised context. Sensitivity analysis and neural networks (NN) enable to representative evaluation of fitness values. About the same time, Kim, et al.[5], used k-means clustering and Expectation Maximization (EM) as embedded unsupervised approach to evaluate a feature subset encoded in chromosomes. The MOEA employed in this case is called evolutionary local search algorithm (ELSA). With these results as a starting point, research of unsupervised learning in feature selection was expanded. Morita, et al.[6] used the k-means clustering algorithm in a wrapper approach, that encoded with Non-dominated Sorting Genetic Algorithm, type two (NSGA-II). Moreover, two objective functions, i.e. the number of features in a set, and a clustering validation (e.g. Davies-Bouldin (DB)[7]) index are introduced. Handl and Knowles[8] examined different combinations of objective functions and Mierswa[1] investigated different indices, i.e. the normalized DB index. More recent work[9] stated that their multi-objective unsupervised feature selection algorithm (MOUFSA) outperforms several other multi-objective and conventional single-objective methods, by using redundant measurements and negative epsilon-dominance. In addition, three new mutation methods are designed to enhance MOUFSA.

However, the defined criteria in classical objective functions used in unsupervised MOEA, fail to predict the performance of clustering results, i.e. the overlapping information (features) in-between classes which probably highlights the essentials that are shared within these classes. To solve this problem, we employ fuzzy criteria in a hybrid filter-wrapper approach. Pioneered by Zadeh[10], fuzzy logic-based systems have been successfully utilized to various application areas, e.g. control system and pattern classification[11]. The comprehensibility of fuzzy criteria, namely the linguistic interpretability of fuzzy partitions and the simplicity of fuzzy if-then rules[12], makes it a promising method to access qualified optimization in MOEA when employed into unsupervised learning. Although fuzzy criteria are addressed in a supervised manner[13], it rarely has been reported in unsupervised cases, in which the natural patterns are discovered according to fuzzy clustering validity and fuzzy objective functions.

In this paper, FS procedure is optimized using the generic heuristic search algorithm NSGA-II, and fuzzy criteria are employed in both filter and wrapper approaches. In the unsupervised learning procedure a new fuzzy index is specifically proposed as one of the objective functions. The target functions are: (i) value of Correlation Membership Measurement (CMM); and (ii) cardinality of feature subset. Here we intend to contribute to the further development of the hybrid methodology, by realizing a sensible integration of fuzzy criteria and MOEA approach in FS area This methodology is applied to a wide set of benchmark datasets and it is compared with commonly used approaches to show its general applicability and competitive advantages.

The remainder of this paper is organized as follows. In Section 2, we introduce the methodology including application of fuzzy criteria and fuzzy model in FS; subsequently, the utilization of NSGA-II in an unsupervised context is presented. In Section 3 experimental results are given and Section 4 conclusions are presented.

2. Methodology

2.1. Fuzzy entropy in filter-approach

In information theory, entropy is a measure of chaos or uncertainty associated with the variables. The concept of entropy has been defined in various ways and used in different fields; fuzzy logic is becoming commonly used in the estimation of entropies. On this basis, we propose an approach embedding fuzzy c-means (FCM)[14] clustering algorithm to estimate the fuzzy entropy by automatically computing the feature memberships. To depict the level of similarity, the feature membership index assigned with a fuzziness characteristic that can be expressed as u_{ij}. In this manner, according to De Luca and Termini[15], the fuzzy entropy can be defined as:

$$H\big(u_j(x)\big) = \frac{1}{n \ln 2} \sum_{j=1}^{n} -u_j(x)\ln u_j(x) - (1-u_j(x))\ln(1-u_j(x)) \quad (1)$$

In Eq. (1), $u_j(x)$ denotes the membership index of the j^{th} feature in the feature pattern vector, meaning every individual feature entropy is computed along all the samples *x*. Subsequently, the entire set of features is ranked for guiding the optimization procedure in the wrapping approach, via maximizing their corresponding fuzzy entropies.

2.2. Fuzzy cost function in wrapping-approach

Multi-objective function optimization, by means of a wrapper technique for unsupervised feature selection, relies on the use of an internal technique of cluster validation. In other words, clustering validation techniques have been designed specifically for the selection of the best clustering solution on the basis of its distance performance. Sometimes the clustering performance is estimated by considering the ratios between intra-class compactness, and inter-class separation. As reported Handl and Knowles[8], this generally suffers from the bias of these measurements with respect to the dimensionality of the feature space. The conflict of this bias can be noticed when dimensionality of a given dataset is enlarged: i.e. the mean of the distribution tends to increase while simultaneously the variance of the distribution decreases. This will cause such a validation technique to be unable to sensitively estimate the difference between all pairs of points, especially in a high dimensional dataset.

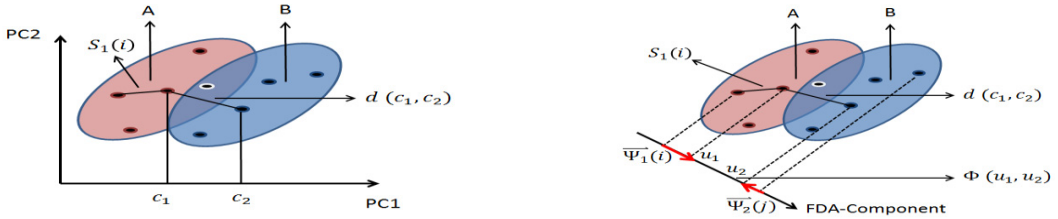

Fig.1. Sketch diagram for CMM. (a) In a high dimensional feature space, two overlapping classes (A and B) with centroids c_1 and c_2 are projected onto two principle component axis PC1 and PC2. *S* is the distance between objects and its belonging center; *d* is the in-between cluster center distance. (b) After Fisher discriminant analysis (FDA) linear projection (project onto the FDA-Component axis), the corresponding components can be rewritten as $d\ (c_1, c_2) \mapsto \Phi(u_1, u_2)$ and $S \longmapsto \overrightarrow{\Psi}$ respectively.

To tackle this bias, we propose a fuzzy cost function, the correlation membership measurement (CMM). This function employs both individual clustering information and shared (overlapping) information (cf. Fig. 1 (a)). We measure the similarity between pairs of vectors using their scalar distance and their directions in high-dimensional attribute space are compared via the projection onto low-dimensional space (cf. Fig. 1 (b)). This is defined as:

$$\mathrm{CMM} = U_{A\cup B} + U_{A\cap B} \tag{2}$$

subject to

$$\begin{cases} U_{A\cup B} = \frac{\frac{1}{N}\sum_{i=1}^{N} S_1(i) + \frac{1}{M}\sum_{j=1}^{M} S_2(j)}{d\,(c_1,c_2)} \\ U_{A\cap B} = \frac{1}{NM}\sum_N \sum_M \{\frac{|\overrightarrow{\Psi_1}(i)-\overrightarrow{\Psi_2}(i)|\cdot|\overrightarrow{\Psi_1}(j)-\overrightarrow{\Psi_2}(j)|}{||\Phi(u_1,u_2)||^2}\} \end{cases}, \quad i \in \text{A and } j \in \text{B} \tag{3}$$

Where in the first term of Eq. (2), i.e. the dependent membership $U_{A\cup B}$ of class A and class B are measured, $S_1(i)$ and $S_2(j)$ are the distance of the vector i and j to their corresponding centroid c_1 and c_2; while $d\,(c_1,c_2)$ is the distance between two cluster centroids, and $\|.\|$ is distant norm as well. N and M are the numbers of the elements that belong to their classes. The evaluation of performance for the overlapping clusters can be achieved by estimating the positions of every individual vectors in a feature subspace. In a high-dimensional domain, however, the comparison of vectors in terms of directions and angles is not applicable. Therefore, principle projection in FDA[16] is used to find a linear combination of features that characterizes two or more classes. The projection matrix can be defined as:

$$\omega = S_w^{-1}(c_1 - c_2) \tag{4}$$

Where

$$S_w = (i - c_1)(i - c_1)^T + (j - c_2)(j - c_2)^T \tag{5}$$

Subsequently, in the second term of Eq. (2), i.e. in the correlated membership $U_{A\cap B}$, the projected vector Ψ and Φ can be obtained by multiplying S norm and d norm with FDA projection matrix ω respectively. Moreover, one should realize that, when applied on a real dataset, the S_w, i.e. the with-in class scatter matrix, normally is a singular matrix and thus non-invertible. We have added a tiny perturbation factor to prevent the projection program from being trapped and the projection matrix is rewritten as:

$$\omega = (S_w + \varepsilon I)^{-1}(c_1 - c_2) \tag{6}$$

Here, I is a unit diagonal matrix. The objective is to achieve proper clustering by minimizing the CMM index. With respect to the aim of feature selection, it is more efficient and direct to use the cardinality of feature subsets as a second cost function. However, one can observe (cf. Fig. 2) that the CMM value decreases with increasing feature numbers. Therefore, a constraint is that at least one feature count in the second objective function should be set.

2.3. Bi-objective optimization

In the previous section, two objective functions (the cardinality of feature subsets and Eq. 4) are formulated as quality indicators for the feature extraction procedure. Those two objective functions are conflicting and form a combinatorial bi-objective optimization problem. Therefore, we aim at searching for the Pareto front, which represents the non-dominated solutions of the proposed feature selection procedure and which can be used to assess the trade-off. In order to achieve this, Evolutionary Multi-objective Optimization Algorithm (EMOA) is adopted due to its capability of handling combinatorial problems. We specifically utilized the well-known NSGA-II[21] algorithm (Non-dominated Sorting Genetic Algorithm) which is the multi-objective extension to the classical Genetic Algorithm[22]. NSGA-II has the ability to generate well-spread Pareto fronts with relatively low computational overhead and it is proved to be robust in real-world applications through numerous testing and applications. In this paper, we omit the detailed discussion on the optimization procedure and use NSGA-II as a 'standard' multi-objective optimizer.

As we are dealing with combinatorial optimization problem, discrete Pareto fronts are obtained from NSGA-II, in

which each point on the resulting Pareto front represents a candidate feature subset. Each candidate solution will be used for the clustering algorithm and the one giving the best clustering performance (cf. the performance indicators in Section 3.2) is chosen. Note that the functionality of the bi-objective optimization is to prescreen the 'bad' candidate solutions (Pareto dominated feature subsets) from all the possible solutions, leaving the Pareto optimal candidates, the number of which is very small compared to the entire number of solution candidates, to be tested in clustering.

3. Experimental results

The objective of this section is to assess the performance of integrating fuzzy criteria into unsupervised multi-objective feature selection procedure. Acceptable results in terms of developing either searching optimization or clustering validation algorithms has been reported in a number of papers. However, for a fair and effective validation of the proposed FC-MOFS method, a commonly used approach without fuzzy constraint[8], referred to as NF-MUFS, is used. Additionally, all datasets are employed in Baseline, using the full feature set. The experiments are conducted on six publicly available datasets, representing multiple disciplines and real life problems (cf. Table 1).

Table 1. Dataset description.

Dataset	Type	Size	Dimension	Class
Glass		214	9	6
Wine	Numerical	178	13	3
WDBC	data	569	30	2
Libras		270	90	15
Sonar	Voice	208	60	2
UMIST	Image	575	644	20

3.1. Parameter setting

In both FC-MOFS and NF-MUFS, the maximum generation and population size are set as same to 100 and 25 respectively; the crossover percentage is 0.9 and the mutation percentage is 0.4, while the rate of mutation is adaptively selected according to the non-dominated sorting performance and expected number of local optima. The clustering algorithm in unsupervised learning of FC-MOFS is fuzzy c-means, which is substituted by k-means in NF-MUFS.

3.2. Validation of FS approach

From the literature, three widely used evaluation metrics, i.e., Accuracy[17] (ACC), Normalized Mutual Information[18] (NMI) and Rand Index[19] (RI) are computed for our experiments. To gain insight in the proposed method, we investigated some aspects that influence clustering performance after feature selection schemes. In the filter approach, the fuzzy entropy feature selection runs once to rank all features for guiding the process in NSGA-II algorithm as initialization; then the results of 20 independent runs of NSGA II to obtain global non-dominated features (cf. Fig. 2) set are tested on six different benchmarks (cf. Table 2 to Table 4). Setting three different evaluation strategies, i.e., the application on full sample population (f-s), random sampling (r-s) on the basis of bootstrapping, and uniform distribution sampling (u-s), the accuracy and general capability of FC-MOFSA are measured in overall 50 times.

The results of bi-objective optimization are illustrated in Fig. 2, in which each subfigure stands for one data set. The blue crosses in the figure represent different candidate feature subsets after the termination of NSGA-II optimizer. Because of the stochasticity of the NSGA-II optimizer, 20 independent runs are conducted for each data set, resulting in a 'layering structure' of the blue crosses. From all the independent runs, we only selected the non-dominated ones using the non-dominated sorting technique. The Pareto fronts generated from 20 independent runs are marked by red circles in Fig. 2. Most of the Pareto fronts are convex, except for Fig. 2(a), in which only 3

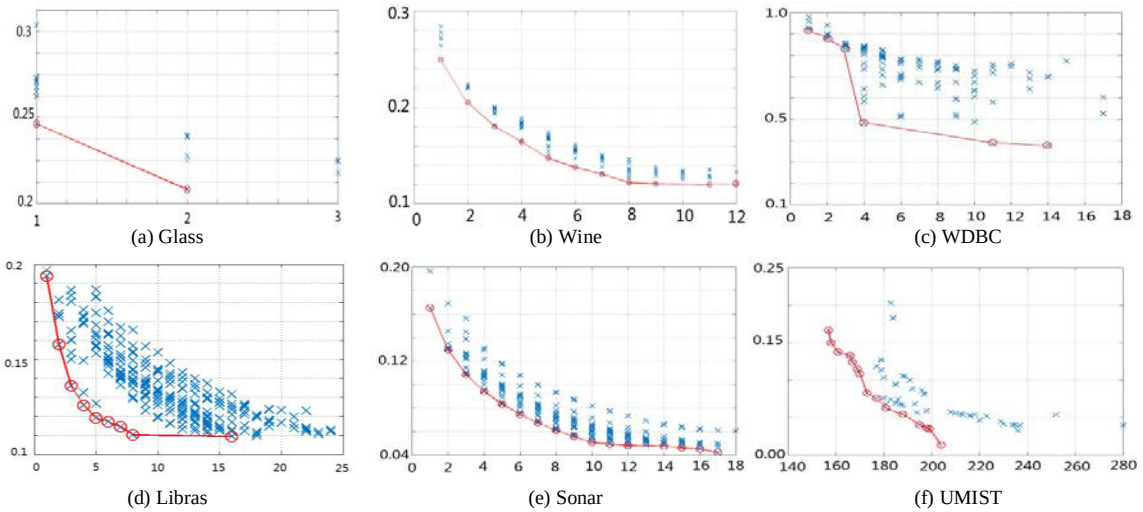

Fig. 2. The Pareto fronts for all dataset ((a) to (f)), consisting of 20 independent runs for each database, including 100 generations per run; the global non-dominated sets are selected (red circle) from local non-dominated sets (blue cross). The vertical axis is CMM error w.r.t the number of features on the horizontal axis.

Table 2. Impact of fuzzy and non-fuzzy feature selection algorithms to the clustering results in ACC index (best-row performance is marked as bold italic).

dataset	Sampling strategy	ACC ± std (%)				
		FC-MOFSA	*nf*	NF-MUFS	*nf*	Baseline
Glass	f-s	**53.93** ± 2.11	2	44.72 ± 2.12	2	44.20 ± 4.15
	r-s	**54.36** ± 3.77	2	42.61 ± 3.39	3	41.76 ± 4.43
	u-s	**55.50** ± 4.09	2	43.36 ± 2.47	1	41.85 ± 4.81
Wine	f-s	**80.34** ± 3.15	2	75.28 ± 3.22	3	70.22 ± 2.28
	r-s	**80.27** ± 5.15	2	75.20 ± 3.45	10	69.02 ± 5.21
	u-s	**78.31** ± 6.34	2	74.58 ± 4.97	3	68.00 ± 7.39
WDBC	f-s	**88.40** ± 2.38	2	83.83 ± 1.85	14	85.41 ± 2.49
	r-s	**88.27** ± 2.15	2	84.38 ± 1.95	14	84.51 ± 2.11
	u-s	**88.37** ± 2.54	3	84.83 ± 2.34	6	84.53 ± 2.64
Libras	f-s	**47.79** ± 4.44	16	44.44 ± 4.29	20	44.81 ± 2.21
	r-s	**28.85** ± 3.35	16	27.67 ± 4.30	20	17.23 ± 2.01
	u-s	28.46 ± 4.22	16	**28.55** ± 4.31	29	17.93 ± 2.28
Sonar	f-s	**57.44** ± 2.99	15	51.44 ± 2.46	4	55.29 ± 3.85
	r-s	**59.67** ± 2.89	5	54.35 ± 2.30	14	55.20 ± 3.73
	u-s	**60.67** ± 3.34	4	54.46 ± 2.70	16	56.37 ± 4.03
UMIST	f-s	**47.91** ± 4.11	167	45.78 ± 2.88	197	43.65 ± 1.48
	r-s	**25.78** ± 2.39	199	22.50 ± 2.48	197	13.43 ± 1.53
	u-s	**25.56** ± 3.20	204	23.33 ± 3.05	197	13.78 ± 1.45

features are present and which indicates the existence of trade-off solutions. In addition, the points on the Pareto front are well-spread. In Fig. 2(c), the distribution of the points is not as good as the rest, which suggests that using more evaluation budget in the multi-objective optimization might improve the quality of the Pareto front on the WDBC dataset. On the basis of our candidate solutions, the resulting Pareto fronts are reliable for using later in the clustering algorithm. The details of six datasets are shown in Table 1. The results of comparisons of clustering performance are listed in Table 2, Table 3 and Table 4. The values indicated in bold are the best results among the algorithms in the same situation and *nf* denotes the number of features used in the clustering. These results suggest the following evaluations: (1) Compared with the baseline, it can be observed that the feature selection procedure is

Table 3. Impact of fuzzy and non-fuzzy feature selection algorithms to the clustering results in NMI index (best-row performance is marked as bold italic).

dataset	Sampling strategy	NMI ± std (%)				
		FC-MOFSA	*nf*	NF-MUFS	*nf*	Baseline
Glass	f-s	**41.25** ± 4.35	2	33.12 ± 2.63	2	39.37 ± 5.42
	r-s	**45.14** ± 4.25	2	35.14 ± 2.93	2	38.60 ± 5.08
	u-s	**47.01** ± 3.33	2	36.62 ± 3.76	2	39.11 ± 5.50
Wine	f-s	**52.37** ± 5.43	2	41.63 ± 5.22	10	42.87 ± 5.19
	r-s	**53.36** ± 0.64	2	44.60 ± 4.86	10	44.95 ± 6.40
	u-s	**52.09** ± 8.40	2	44.09 ± 5.61	7	44.95 ± 7.90
WDBC	f-s	**44.79** ± 5.35	1	38.02 ± 4.28	28	42.20 ± 5.08
	r-s	**41.17** ± 5.01	1	38.56 ± 4.44	4	40.41 ± 4.32
	u-s	39.85 ± 5.45	2	39.90 ± 5.24	6	**41.42** ± 5.25
Libras	f-s	**62.10** ± 2.83	16	56.36 ± 3.33	29	60.84 ± 3.44
	r-s	**25.98** ± 3.00	16	20.80 ± 3.24	16	19.93 ± 3.39
	u-s	**28. 85** ± 2.59	16	22.67 ± 3.39	29	22.01 ± 3.52
Sonar	f-s	**0.91** ± 0.81	14	0.91 ± 1.83	4	0.88 ± 0.87
	r-s	**2.53** ± 0.71	5	1.82 ± 0.79	14	1.21 ± 0.73
	u-s	**2.84** ± 1.11	4	1.95 ± 1.47	16	1.64 ± 0.81
UMIST	f-s	63.84 ± 4.04	167	**64.74** ± 4.83	167	63.82 ± 1.83
	r-s	**25.57** ± 3.95	199	20.17 ± 3.86	197	13.10 ± 2.12
	u-s	**28.39** ± 4.67	204	22.43 ± 4.98	197	14.86 ± 1.63

Table 4. Impact of fuzzy and non-fuzzy feature selection algorithms to the clustering results in RI index (best-row performance is marked as bold italic).

dataset	Sampling strategy	RI ± std (%)				
		FC-MOFSA	*nf*	NF-MUFS	*nf*	Baseline
Glass	f-s	**65.49** ± 2.15	2	58.94 ± 2.88	2	53.63 ± 4.32
	r-s	**65.97** ± 2.17	2	58.22 ±2.62	2	48.89 ± 3.60
	u-s	**65.59** ± 2.03	2	58.35 ± 2.93	2	44.35 ± 1.58
Wine	f-s	**77.86** ± 3.02	1	73.00 ± 2.99	3	71.86 ± 5.58
	r-s	**78.03** ± 3.90	1	74.53 ± 3.20	3	43.66 ± 5.02
	u-s	**76.48** ± 4.90	1	74.01 ± 4.10	3	44.91 ± 5.46
WDBC	f-s	73.79 ± 3.58	1	73.08 ± 2.99	5	**75.04** ± 2.19
	r-s	**74.34** ± 3.02	1	73.64 ± 2.68	14	50.70 ± 5.51
	u-s	**74.46** ± 3.89	2	74.27 ± 3.26	6	50.46 ± 5.92
Libras	f-s	**90.40** ± 4.85	16	90.16 ± 3.25	20	90.37 ± 7.85
	r-s	90.68 ± 4.76	16	**91.30** ± 4.66	29	83.87 ± 7.87
	u-s	**91.55** ± 4.95	16	91.29 ± 6.26	18	82.34 ± 2.45
Sonar	f-s	**50.80** ± 6.55	4	49.70 ± 6.88	4	50.32 ± 4.19
	r-s	**51.11** ± 5.98	5	50.16 ± 5.08	4	49.97 ± 3.91
	u-s	**51.18** ± 8.53	4	50.14 ± 8.78	4	49.99 ± 6.42
UMIST	f-s	**95.51** ± 6.11	198	88.51 ± 4.37	197	92.80 ± 1.48
	r-s	**94.69** ± 5.24	199	86.95 ± 4.58	167	88.01 ± 1.04
	u-s	**94.40** ± 6.65	199	89.11 ± 5.12	167	85.72 ± 1.24

necessary and efficient by removal of noise and redundancy . (2) The best solutions of the proposed FC-MOFSA mostly have higher accuracy, mutual information and RI other than the non-fuzzified feature selection algorithms (NF-MUFS and Baseline employment). In spite of the slightly less performance on WDBC, Libras and UMIST dataset, the u-s and f-s value are still competitive compared with the best results of other methods. (3) The average r-s means that even though with less samples (information) obtained from entire population, still, in most situations, the results of FC-MOFSA are better than those of NF-MUFS and Baseline. (4) The proposed method, in most cases, has the least numbers of features for prediction of the best results. In the second highest cases, FC-MOFSA still

obtains the lowest cardinality of feature sets. (5) By expressing the descriptor of similarity in RI and descriptor of redundancy in NMI, our method achieves an accurate clustering performance. This is due to the exploitation of discriminative and overlapping information in an unsupervised context. (6) The accuracy and the similarity grouping capability of the experimental algorithms suffer from a serious degradation when down-sampling is applied on the Libras and UMIST dataset. The sparse distribution of these dataset complicates the unsupervised categorization scheme. However, it is observed that FC-MOFSA is superior to the rest approaches by uncovering the underlying patterns and possibly skewed structure.

4. Conclusions

In this paper, we present a new multi-objective feature selection algorithm utilizing the fuzzy hybrid filter-wrapper approach. We introduce a fuzzy criterion-based manner in multi-objective optimization problems and thereby increase the clustering accuracy in unsupervised feature selection schemas. The proposed method outperforms the commonly used multi-objective feature selection method with non-fuzzified parameters, in terms of accuracy and general capability. In addition to the fuzzy entropy in pre-selection, we also present a new fuzzy index called Correlation Membership Measurement (CMM), which produces superior results, particularly on sparse and skewed datasets. Future work will focus on further comprehensive and systematic validation considering different combinations of clustering algorithms and objective functions[3] using principles of fuzziness.

References

1. Mierswa I, Wurst M. Information preserving multi-objective feature selection for unsupervised learning. *Proceedings of the 8th annual conference on Genetic and evolutionary computation*, pp. 1545-1552. ACM, 2006.
2. Li Z, Lu H. Clustering-guided sparse structural learning for unsupervised feature selection. *IEEE Transactions on Knowledge and Data Engineering*. 2014 Sep; 26(9):2138-50.
3. Mukhopadhyay A, Coello CA. A survey of multiobjective evolutionary algorithms for data mining: Part I. *IEEE Transactions on Evolutionary Computation*. 2014 Feb; 18(1):4-19.
4. Oliveira LS, Suen CY. Feature selection using multi-objective genetic algorithms for handwritten digit recognition. In *Pattern Recognition, 2002. Proceedings*. 16th International Conference on 2002 Vol. 1, pp. 568-571. IEEE.
5. Kim Y, Menczer F. Evolutionary model selection in unsupervised learning. *Intelligent data analysis*. 2002 Jan 1; 6(6):531-56.
6. Morita ME, Suen CY. Unsupervised Feature Selection Using Multi-Objective Genetic Algorithms for Handwritten Word Recognition. In *ICDAR* 2003 Aug 3 Vol. 2, pp. 666-670.
7. Davies DL, Bouldin DW. A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*. 1979 Apr; (2):224-7.
8. Handl J, Knowles J. Feature subset selection in unsupervised learning via multiobjective optimization. *International Journal of Computational Intelligence Research*. 2006 Jun; 2(3):217-38.
9. Xia H, Yu D. Multi-objective unsupervised feature selection algorithm utilizing redundancy measure and negative epsilon-dominance for fault diagnosis. *Neurocomputing*. 2014 Dec 25;146:113-24.
10. Zadeh LA. Fuzzy sets as a basis for a theory of possibility. *Fuzzy sets and systems*. 1999 Dec 31;100:9-34.
11. Lajoie SP, Derry SJ, editors. Computers as cognitive tools. *Routledge*; 2013 May 13.
12. Ishibuchi H, Yamamoto T. Fuzzy rule selection by multi-objective genetic local search algorithms and rule evaluation measures in data mining. *Fuzzy sets and systems*. 2004 Jan 1;141(1):59-88.
13. Vieira SM, Kaymak U. Fuzzy criteria for feature selection. *Fuzzy Sets and Systems*. 2012 Feb 16;189(1):1-8.
14. Trivedi MM, Bezdek JC. Low-level segmentation of aerial images with fuzzy clustering. I*EEE Transactions on Systems, Man, and Cybernetics*. 1986 Jul;16(4):589-98.
15. De Luca A, Termini S. A definition of a nonprobabilistic entropy in the setting of fuzzy sets theory. *Information and control*. 1972 May 31;20(4):301-12.
16. Scholkopft B, Mullert KR. Fisher discriminant analysis with kernels. *Neural networks for signal processing* IX. 1999 Aug;1(1):1.
17. Papadimitriou CH, Steiglitz K. Combinatorial optimization: algorithms and complexity. *Courier Corporation*; 1982.
18. Ghosh J, Acharya A. Cluster ensembles. *Wiley Interdisciplinary Reviews*: *Data Mining and Knowledge Discovery*. 2011 Jul 1;1(4):305-15..
19. Rand WM. Objective criteria for the evaluation of clustering methods. J*ournal of the American Statistical association*. 1971 Dec 1;66(336):846-50.
20. Zitzler E, Thiele L. Multiobjective evolutionary algorithms: a comparative case study and the strength Pareto approach. I*EEE transactions on Evolutionary Computation*. 1999 Nov;3(4):257-71.
21. Deb K, Meyarivan TA. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE transactions on evolutionary computation*. 2002 Apr;6(2):182-97.
22. Holland JH. Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence. *U Michigan Press*; 1975.