Modulated Clustering Using Integrated Rough Sets and Scatter Search Attribute Reduction

Abdel-Rahman Hedar*
Dept. of Computer Science
Faculty of Comp. & Informatics
Assiut University
Assiut 71526, Egypt
hedar@aun.edu.eg

Abdel-Monem M. Ibrahim
Dept. of Mathematics
Faculty of Science
Al-Azhar University
Assiut 71524, Egypt
abdelmonem@azhar.edu.eg

Alaa E. Abdel-Hakim§
Electrical Engineering
Department
Faculty of Engineering
Assiut University
Assiut 71516, Egypt
alaa.aly@eng.au.edu.eg

Adel A. Sewisy
Dept. of Computer Science
Faculty of Comp. & Informatics
Assiut University
Assiut 71526, Egypt
sewisy@aun.edu.eg

ABSTRACT

Feature selection is an essential problem in pattern classification systems. The entire performance of the classifier is highly affected by the quality of the selected features. In this paper, we address this problem by integrating feature selection with the clustering process. A novel feature-modulated classification algorithm is proposed to improve the classification accuracy. We use a rough sets approach for feature selection based on a scatter search meta-heuristic scheme. The proposed approach sifts a compact subset of characterizing features in multi-class systems according to the clustering performance. To verify the effectiveness of our method, experimental comparisons are carried out on eleven benchmark datasets using two typical classifiers. The results indicate that the proposed method has a remarkable ability to generate effective reduced-size subsets of salient features while yielding significant classification accuracy.

Keywords

Unsupervised classification; data clustering; attribute reduction; scatter search; rough sets.

1. INTRODUCTION

Data mining techniques are usually used to understand existing patterns in the data and to generate predictive or statistical models that could describe the data behavior and its changes [3, 5]. While data mining includes an extensive variety of data processing and manipulation problems, machine learning methods are essential tools in data mining [20].

Feature selection or attribute reduction is a standout amongst major problems in the field of machine learning. The fundamental purpose behind feature selection is to decide which minimal feature subset should be selected in order to retain a high accuracy in representing the whole feature set. So, feature selection ends up being the way toward finding a subset of condition features that is necessary and sufficient to describe the decision feature or target related to the original set of features in a given information system [36].

One of the most important requirements for feature selection is to develop smart methods with abilities to handle imprecise and inconsistent information such as noisy, irrelevant, or misleading features [25, 26]. Rough set theory is a powerful concept that could handle uncertainty, vagueness, and inconsistent data. Based on rough set concepts, the feature selection or attributes reduction problem can be formulated to predict decision concepts based on a subset of features or attributes similar to that decision obtained by the original feature set [35, 28]. Feature selection algorithms vary in terms of their sophistication level. One can build a simple feature selection mechanism by collecting the common strongly relevant features in addition to some other weakly relevant features. Using mutual information and discernibility matrix are very common concepts in feature selection [33, 37]. Moreover, advanced mechanism were built using global search methods or metaheuristics to solve such problem, such as tabu search, simulated annealing, genetic algorithm, ant colony optimization, and others [1, 11, 29, 30].

Data classification is another main problem of data mining. Several real world applications are formulated or trans-

*A. Hedar is also affiliated in the Dept. of Computer Science, Jamoum University College, Umm Al-Qura University, Makkah, Saudi Arabia. Email: ahahmed@uqu.edu.sa

§A. E. Abdel-Hakim is in a leave to the Dept. of Computer Science, Jamoum University College, Umm Al-Qura University, Makkah, Saudi Arabia. Email: adali@uqu.edu.sa

GECCO '18 Companion, July 15–19, 2018, Kyoto, Japan

DOI: https://doi.org/10.1145/3205651.3208286

ferred to classification of giving data such as document classification, speech recognition, object classification in computer vision, etc. [19]. Therefore, there is a great need to develop powerful and accurate classifiers to fit the growth of such classification problems especially for applications that need quick real-time response. To solve the general data classification problem, three main steps are implemented. The first step is to select a training dataset with class labels. Then, a model or description for each class is developed by studying the relations between the attributes of the data elements in each class. Finally, the prediction quality of the obtained model is evaluated. Several theoretical and heuristic methods have been developed in order to enhance the quality of classification models [22, 21].

The size of data objects and/or data features is not usually large in several applications. The process of collecting such application data does not consider data mining as the main purpose behind collecting that data. Therefore, there are high possibilities of finding data redundancy or feature irrelevancy. Hence, searching the sample space is needed in order to extract the valuable information from the given data. Searching such spaces usually suffers from high computational time complexity. A practical remedy is to reduce the search space dimensionality and predict compact models. In practical classification applications, the training data usually contain large numbers of features and a limited number of sample objects. Examples of such cases are spam email classification, URLs ranking, gene selection from microarray data, etc. [34, 10]. Actually, the classification accuracy can be improved by applying an effective feature reduction [14]. Moreover, the classification processing time can be reduced if we could use less number of features which representing the whole feature set.

In this work, we propose a Scatter Search Rough Set Attribute Reduction (SSAR) algorithm for the feature subset selection. Then, we investigate the effect of SSAR on improvement of the classification accuracy of two clustering algorithms: Simulated Annealing combined with ellipsoidal clustering technique (AELC) [13], and K-Means Cloning (KMC) algorithm presented in [12]. Furthermore, we compare our approaches with two other algorithms, Artificial Bee Colony (ABC) algorithm [16] and Particle Swarm Optimization (PSO) [4]. Eleven benchmark datasets are used for testing and evaluation purposes in order to demonstrate the effectiveness of the proposed methodology.

The rest of the paper is structured as follows: in Section 2, we briefly review the principles of rough set and scatter search meta-heuristic. In Sections 3, we explain the main components of the proposed modulated-classification method. Then, the experiments and results are presented and discussed in Sections 4 and 5, respectively. The conclusion and remarks for future work are provided in Section 6.

2. PRELIMINARY TECHNIQUES

In the rough set framework, an attribute subset is called a reduct if it has a minimal size and the same distinction power as all of the entire attribute set. Obtaining reducts of an information system is a one of the main problems in the rough set theory [24, 25, 15, 32]. Calculating reducts of an information system is needed to extract rule-like knowledge. Clearly, attribute reducts should maintain minimal redundancy as well as retaining the representational power. There are several attempts in the literature to develop efficient algorithms for estimation of useful reduction of information systems, e.g. [18].

We propose a rough set feature selection algorithm that is based on a search method called Scatter Search (SS). We use SS to select more efficient feature subsets that can optimally model the original feature set for the sake of better accuracy. This reduced feature set is then incorporated in the unsupervised classification process.

2.1 Rough Set Theory

The rough set (RS) theory is one of the data mining techniques that are mainly used for dimensionality reduction of data sets and for discovering hidden patterns and generating decision rules [29, 24, 25]. RS is a system for identifying and recognizing regular patterns in input data, particularly if the given data is incomplete and/or has uncertainty. The establishments of RS depend on the set approximation on classification space. RS looks at objects that are characterized by the same information as indiscernible given the available information about them. The most informative attributes are selected in order to make the minimal subset of reducts using RS. These attributes contain conditional attributes, which are most predictive to class attributes. Therefore, unnecessary and uncertain attributes can be removed from the dataset with minimal information loss.

Assume $S = (U, \mathbb{A})$ is an information system with a non-empty set of objects U and a non-empty set of attributes $\mathbb{A}$. For a subset of attributes $B \subseteq A$, an equivalence relation $IND(B)$ can be computed on the set U as [29, 36]:

$$IND(B) = \{(\xi, \eta) \in U \times U | \forall b \in B b(\xi) = b(\eta)\}.$$

This relation is called the indiscernibility relation and its partitions are denoted by $U/IND(B)$.

Assume that U can partitioned into n classes X_i, $1 \leq i \leq n$, by mean of $IND(B)$. Then, the entropy of an attribute set $B \subseteq A$ can be defined as:

$$H(B) = -\sum_{i=1}^{n} (p(X_i)) log(p(X_i)), \quad (1)$$

where $p(X_i) = |X_i|/|U|$, and $|.|$ is the cardinality.

The conditional entropy of an attribute set $E \subseteq A$ with reference to another attribute set B can be defined as follows:

$$H(E|B) = -\sum_{i=1}^{n} (p(X_i)) \sum_{j=1}^{m} (p(Y_j|X_i)) log(p(Y_j|X_i)), \quad (2)$$

where

$$\begin{aligned} p(Y_j|X_i) &= |Y_j \cap X_i|/|X_i|, \\ U/IND(D) &= \{Y_1, Y_2, \ldots, Y_m\}, \\ U/IND(E) &= \{X_1, X_2, \ldots, X_n\}, \end{aligned}$$

for $1 \leq i \leq n, 1 \leq j \leq m$.

2.2 Scatter Search

Scatter Search (SS) is an evolutionary algorithms which starts with a limited number of solutions. At each iteration, some of the best and diverse solutions are selected for combination. Then, trial points can be generated from this combination process. These processes are reiterated until no

more new solution can be found [9]. The structure of the SS has much flexibility than most of other evolutionary algorithms. Moreover, SS invokes an adaptive memory-based search in order to achieve more efficient global search [17].

Several hard optimization problems could be solved by SS since it can be applied to both combinatorial and continuous optimization problems. The search process on SS takes place in a systematic way unlike to the random designs of other methods. Glover [7, 8] introduced the reference set of solutions and several guidelines. He suggested a five-procedure structure as follows:

1. *Diversification Generation Procedure* to generate population P.
2. *Improvement Procedure* to transform a solution into one or more enhanced solutions.
3. *Reference Set Update Procedure* to build the reference set.
4. *Subset Generation Procedure* to generate subsets of reference set as an initial stage to produce new combined solutions.
5. *Solution Combination Procedure* to produce one or more new combined solution from each subset generated by the Subset Generation Procedure.

3. METHODOLOGY

In this section, we discuss the components of the proposed algorithm. These components can be summarized as follows:

- An attribute reduction technique based on the scatter search for attribute reduction (SSAR). The main target of this technique is to reduce the dimensionality of classification space.
- Data clustering techniques based on modified two versions of K-means methodology called K-mean cloning (KMC) [12] and adaptive ellipsoidal clustering (AELC) [13]. These techniques are used to extract features of each class via the training data.
- An unsupervised classification system to predict the class of a given data instance using the trained classifier that is obtained by the clustering techniques.

3.1 Scatter Search for Attribute Reduction

A feature selection technique, called Scatter Search Rough Set Attribute Reduction (SSAR), has been presented in [36]. The SSAR methods tries to find subsets of attributes that could replace the whole attribute set without losing merits of predicting the decision attribute.

The SSAR starts with a population of solutions which are coded in binary vectors. The dimension of each solution equals the number of conditional attributes $|\mathbb{C}|$. In this coding system, the value 1 for a component i, $i = 1, \ldots, |\mathbb{C}|$, means that the i-th condition attribute is contained in the contained in this solution. Otherwise, the solution excludes this i-th condition attribute. Those solutions are evaluated based on two criteria:

- Their entropy values:
  - The entropy functions given by Equations (1) and (2) are evaluated for the generated solutions in order to estimate their quality. Then, solutions with lower entropy values are favored.
- The cardinality of the condition attribute subsets represented by those solutions:
  - If the compared solutions have the same entropy value, then the solution with lower cardinality are favored.

The SSAR method modifies the main structure of SS, given in Section 2.2, in which six procedures instead of five are used. The new added procedure is:

6. *Intensification Procedure* to refine best solutions found so far.

3.2 Data Clustering by KMC and AELC

A cluster validity method is exploited as an additional criterion to determine the appropriate clusters. The methods proposed adaptive spherical and ellipsoidal clustering techniques combined with simulated annealing, named (KMC) [12] and (AELC) [13], respectively. AELC expands the scope from dealing with data involving certain solid patterns, such as spherical clusters as in KMC, to handle other categories containing ellipsoidal clusters and non-symmetric forms. Silhouette Function (SF) [27] is known as the best for automatic clustering analysis, especially in determining the optimal clusters of the type of circular or spherical clusters. This should be effective in the case that the number of clusters is not known. KMC and AELC have improved the performance of the silhouette width-based objective function in combination with the nearest-neighbor algorithm to be more extensive by giving priority to the appropriate/optimal number of localities in ellipsoidal clusters. So, AELC presented a new method for assignment of data instances based on elliptical rules, instead of relying on K-mean, which fails to cluster data instances that have ellipsoidal patterns.

The main idea of the proposed algorithms is to utilize simulated annealing heuristics to generate non-local trials for the cluster centers in order to find better solutions. Simulated annealing is used with KMC and AELC algorithms while retaining all the previous steps of SSAR. The procedure is shown as follows:

1. *Initialization:* Choose K centroids randomly as initial centers, the cooling schedule parameters: initial temperature, final temperature and cooling rate.
2. *Generation:* A trial solution selects a neighbor solution and calculates the corresponding data clustering quality.
3. *Clusters Determination Procedure:* The crucial step is how to optimally select the next cluster(s) for splitting or merging. Split and merge by SF issues criterion functions to choose clusters to be split or merged and evaluate the cluster structure changes.
4. *Splitting / Merging Procedures:* New clusters are created by either merging two smaller clusters into a bigger one or splitting a larger cluster into two or more smaller ones.

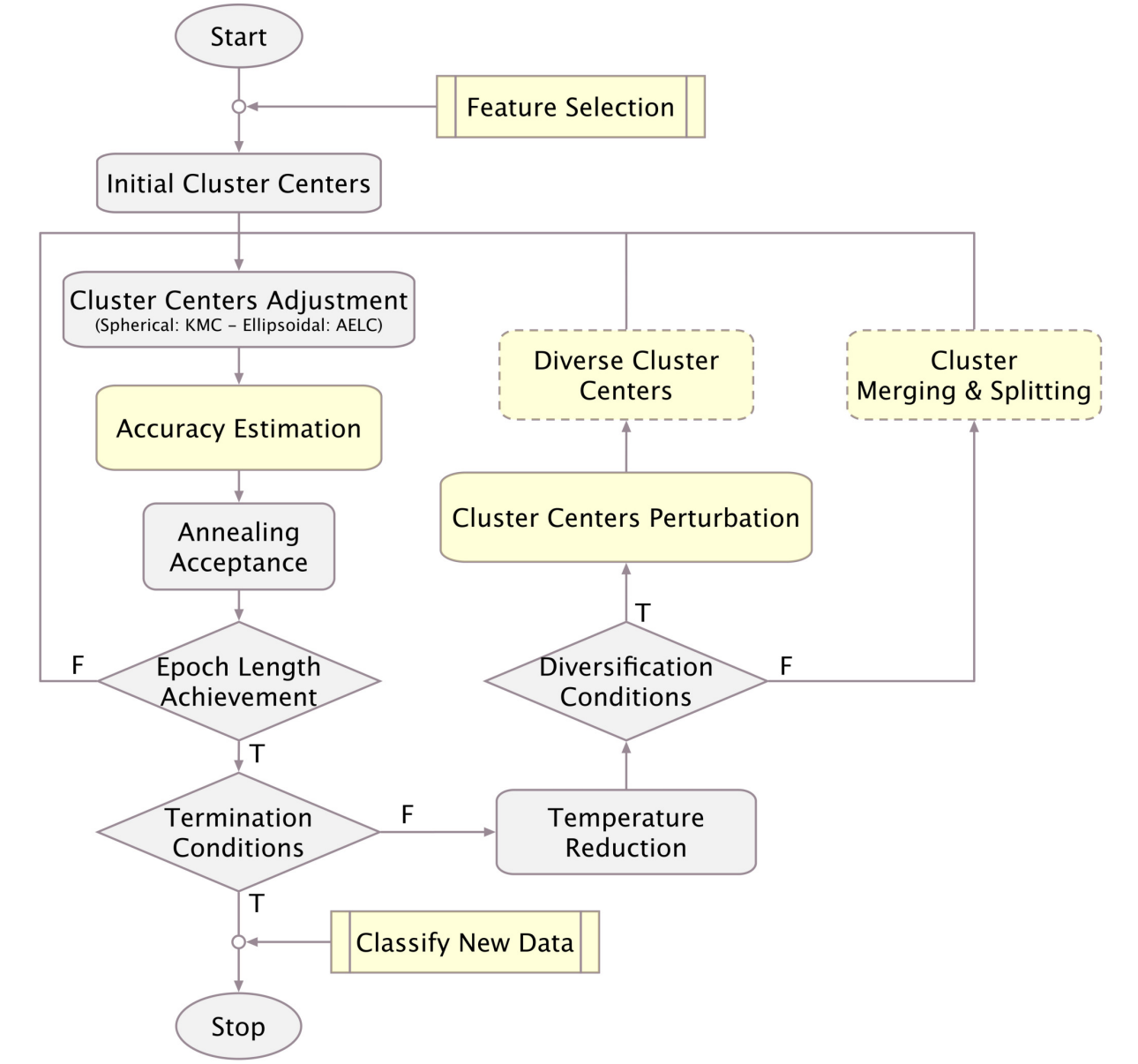

Figure 1: The flowchart of the proposed method

At each stage, where new solutions are to be generated, the method of assignment of points will be applied. KMC and AELC can determine the optimal number of clusters after improving the search method and point assignment to the best given clusters. Moreover, AELC is expandable to different data forms, especially the ellipse clusters.

3.3 Data Classification

Clustering is a technique for unsupervised learning in which observations are gathered to be in similar clusters in some sense. In this paper, a new modulation classification method based on the combination of feature selection and clustering, in which a new algorithm is introduced to extract key features. KMC and AELC clustering are modified and adopted to recover cluster forms within variable number of clusters. Hence, a cluster validity measure is calculated to extract key features, which discriminate different modulation types.

The main hypothesis can be summarized as follows: If the main classification algorithm is applied at the first stage on a full features/attributes dataset, the performance will be degraded. This is because feature selection and classification techniques have been applied independently without considering the interaction between them. Therefore, SSAR will be applied to extract reducts before the classification procedure for improved performance.

SSAR and AELC algorithms have been modified to deal with classification problem. The main algorithm is shown in Figure 1 consists of three major components:

1. Calling SSAR to find a minimal subset of features (in case of using classification of salient features only).

2. Calling the KMC or AELC clustering algorithm to predict accuracy of the classifier for the considered features of data.

3. Calculation of the classification accuracy.

The clustering methodology in the proposed framework, which is shown in Fig. 1, introduces the following new elements:

- The classification accuracy is computed in 'Accuracy Estimation' Step after getting the clusters centers. If such accuracy is acceptable, then the steps, with shaded boundary, 'Diverse Cluster Centers' and 'Cluster Merging & Splitting' for changing the cluster number, are disabled.

- Another diversification technique, called 'Cluster Center Perturbation' is used to find new diverse centers without changing the number of clusters.

4. EXPERIMENTAL SETUP

In this section, we describe the experimental setup that was used to evaluate the proposed algorithm. Eleven classification datasets from the well-known UCI database [2] repository, are used to evaluate the classification performance. The datasets are described in Table 1. For comparison purposes, these eleven benchmark are chosen exactly the same as in [16]. 75% of data are randomly picked for training, and the remaining 25% of data are used for testing. The exact sizes of the train and test sets are shown in the last two columns of Table 1.

Table 1: Descriptive statistics of the datasets used in experimentation.

Dataset	Instance	Feature	Class	Train	Test
Balance	625	4	3	469	156
Cancer	569	30	2	427	142
Cancer-Int	699	9	2	524	175
Credit	690	15	2	518	172
Diabetes	768	8	2	576	192
E. coli	327	7	5	245	82
Glass	214	9	6	161	53
Horse	364	26	3	273	91
Iris	150	4	3	112	38
Thyroid	215	5	3	162	53
Wine	178	13	3	133	45

4.1 Datasets

Brief descriptions of the datasets used in the quantitative evaluation experiments as are stated as follows:

The "Balance" dataset was collected to model psychological experimental results. Each instance represents a position of the balance scale tip: to the right, tip to the left, or balanced. The dataset includes 4 attributes, 3 classes and there are 625 instances. Those instances are split into 469 for training and 156 for testing.

The "Cancer" dataset represents "breast cancer Wisconsin - Diagnostic" dataset, while "Cancer-Int" dataset is based on the "breast cancer Wisconsin - Original" dataset. They are diagnosis of breast cancer, with two outputs for tumor categories: either benign or malignant. The first one includes 569 instances, 30 attributes and the second one includes 699 instances and 9 attributes.

The "Credit" (the Australian credit card) dataset is for assessment of credit cards applications. The assessment depended on 14 attributes, including six numeric values and eight discrete ones. There are 690 applicants in total and the output has two classes.

The "Diabetes" dataset is a binary classification dataset to diagnose diabetes as positive or negative. The dataset has 768 instances. We used 576 instances for training and the remaining 192 for testing. There are eight attributes for each instance.

The original dataset of "Escherichia coli" (E. coli) contains 336 instances composing eight classes. Three out of these eight classes contains five or less instances. Therefore, the instances of these three classes, exactly nine instances, are discarded. So, only a total of 327 instances from five classes with seven inputs are considered. 245 out of them are used for training and the remaining 82 instances for testing.

The "Glass" dataset was generated to classify glass types. Those types are vehicle windows, tableware, float processed building windows, non-float processed building windows, head lamps, or containers. Nine attributes are based on chemical measurements determines one of six types of glass including 17, 13, 70, 76, 29, and 9 instances of each class, respectively. A total of the instances is 214 which are split into 161 for training and 53 for testing.

The "Horse" dataset can be used to predict the fate of

a horse with a colic and to classify an expectation of the horse survival, death, or being euthanized. The dataset is generated based on the Horse Colic data with 364 instances and 27 attributes.

The Iris dataset includes 150 objects of flowers from the Iris three species: Setosa, Versicolor, and Virginica. Each of the three classes includes 50 instances with four variables; petal length, petal width, sepal length, and sepal width.

The "Thyroid" dataset contains a diagnosis of thyroid, regardless of whether it is hyper or hypo-function. Five inputs are used to classify three classes of thyroid function as being over function, normal function, or under function. The dataset was collected to contain 215 patterns and 5 features.

The "Wine" dataset was generated from chemical analysis processes of wines that were obtained from three different cultivators. The data analysis processes could distinguish between three types of wines. That dataset contains 178 objects of wine samples with 13 attributes.

5. RESULTS AND DISCUSSION

For each dataset, the value of the Classification Error Percentage (*CEP*) is reported as the percentage of the incorrectly classified instances of the test datasets. Each instance is classified by assigning it to the class whose the closest Euclidean distance center to the center of the clusters. The assigned class is compared with the ground truth and for match/mismatch decision. *CEP* is calculated as the ratio between the number of misclassified and total instances, as shown in the following equation:

$$CEP = 100 \times \frac{P_{vm}}{P_v}$$

where P_{vm} refers to the number of misclassified instances in the test and P_v is the total size of the test dataset.

The classification results are presented in Table 2 for the following clustering algorithms: a) AELC, b) KMC, c) Honey Bees Dancing Language (HBDL) [23], d) Artificial Bee Colony (ABC), which is originally designed by simulating the foraging behavior of honey bees for numerical optimization problems [16], and e) Particle Swarm Optimization (PSO), which mimics the social behavior of a swarm of birds flying to destinations [4]. Table 2 shows the raw classification results without any attribute reduction.

As shown in Table 2, AELC and KMC algorithms outperforms ABC and PSO algorithms in 6 problems, whereas the ABC algorithm results are better than that of AELC algorithm only for one dataset, the Cancer and Credit problems. The same performance is repeated for the other four datasets, the Balance, Cancer-Int, Iris, and Wine problems in terms of classification error. KMC algorithm results are better than those of the ABC algorithm in five datasets, while the latter shows better results in four datasets. Moreover, their results are neutral in terms of classification error for two datasets: the Iris and Wine. Also, AELC algorithm outperforms KMC algorithm in five datasets, whereas KMC algorithm result is better than that of AELC algorithm only for two datasets, the Glass and Horse problems, and the same results are for the other four datasets. Moreover, the average percentages of the classification error for four datasets are identical for AELC and for ABC.

In Table 3, the classification error percentages of AELC and KMC algorithms on eight real-world benchmark classification datasets are presented. The symbols n and n_s denote the total number of the original features and the selected ones, respectively. The averages of the processing time in seconds for running each method are also given in Table 3. These time averages are denoted by T_{AELC} and T_{KMC} for running AELC and KMC without feature selection, and denoted by T^s_{AELC} and T^s_{KMC} in the case of using feature selection, respectively.

Table 2: CEP values of the used classifiers without attribute reduction.

Dataset	AELC	KMC	HBDL	ABC	PSO
Balance	15.38	24.35	**13.56**	15.38	25.47
Cancer	**2.11**	8.45	2.67	2.81	5.80
Cancer-Int	**0**	0.57	1.34	**0**	2.87
Credit	19.18	34.30	**10.07**	13.37	22.96
Diabetes	21.87	21.87	**19.89**	22.39	22.50
E. coli	**9.87**	11.11	13.27	13.41	14.63
Glass	32.07	**26.41**	40.01	41.50	39.05
Horse	30.76	**27.47**	37.63	38.26	40.98
Iris	**0**	**0**	0.34	**0**	2.63
Thyroid	**1.88**	**1.88**	2.74	3.77	5.55
Wine	**0**	**0**	1.28	**0**	2.22

As shown in Table 3, SSAR could find a smaller number of features for different datasets. For example, SSAR selected two features out of a set of 30 features in the cancer dataset. It also selected five features out of a set of 26 features in the horse dataset. Moreover, SSAR could help both methods to severely reduce the processing time.

The positive effect of attribute reduction is obvious, when we look at the classification error percentage. For example, in the horse dataset, AELC produced a CEP of 30.7692% with all features, while CEP is 24.1758% with a reduced feature set of five. Similarly, AELC produced a CEP of 28.3019% with the substantially reduced number of features of two, while CEP is 32.0755% with all of the nine features. Other similar types of scenarios can also be noticed for all the rest of datasets in AELC. Therefore, we may conclude that AELC has powerful searching abilities to obtain high quality solutions. On the other hand, KMC produced small CEP of features reduced for all classes datasets, We can observe that for the performance of AELC and KMC of salient features n_s, the CEP of KMC was better than AELC for glass dataset of 22.6415% when compared to 28.3019% of AELC, whereas AELC was the best of KMC for the rest of the datasets.

The Wilcoxon rank-sum test [6, 31] is invoked to check the statistical significance differences between the obtain results in Table 3. The statistical test results are recorded in Table 4, and the level of significance for these results is 0.05. The statistical tests show the significant reduction in processing time in classification with feature selection compared to classification without feature selection. Moreover, there are no significant differences in classification rates between the two cases.

From these results, we can appreciate the effect of integrating the attribute reduction with classifiers in terms of simplifying the input feature space and achieving better pro-

Table 3: Performance of AELC and KMC for all features and the effect of attribute reduction on the classification accuracy. The symbols n and n_s denote the total number of the original features and the selected ones, respectively. Averages of processing times are provided in seconds.

	Results with all features					Results with selected features				
	n	AELC	KMC	T_{AELC}	T_{KMC}	n_s	AELC	KMC	T^s_{AELC}	T^s_{KMC}
Cancer	30	**2.1127**	8.4507	2163	1723	2	16.9014	16.9014	140	116
Cancer-Int	9	**0**	0.5747	827	647	4	1.7241	2.2989	339	282
Diabetes	8	**21.8750**	**21.8750**	787	617	3	24.4792	27.08	278	231
E. coli	8	9.8765	11.1111	369	280	3	**6.1728**	9.8765	121	100
Glass	9	32.0755	26.4151	270	205	2	28.3019	**22.6415**	55	45
Horse	26	30.7692	27.4725	1186	948	5	**24.1758**	26.3736	220	183
Iris	4	**0**	**0**	99	66	3	**0**	2.7027	57	47
Wine	13	**0**	**0**	322	242	5	**0**	2.3256	107	89

Table 4: Rank-Sum test for the comparison results in Table 3

Comparisons		R^+	R^-	p-value	Best Method
n	n_s	36	0	0.0009	n_s
AELC	AELC/FS	19.5	16.5	0.9596	–
KMC	KMC/FS	9	27	0.6227	–
T_{AELC}	T^s_{AELC}	36	0	0.0148	T^s_{AELC}
T_{KMC}	T^s_{KMC}	36	0	0.0207	T^s_{KMC}

cessing time and classification performance, or at least with acceptable loss in the worst cases of classification.

6. CONCLUSION

In this paper, we proposed an approach to modulate unsupervised clustering process by attribute reduction. Reducing the input feature dimensionality achieves two benefits. The first is the simplification of the input feature space, and consequently the decrease of the computational requirements of the used classifiers. The second benefit is reducing the redundancy of the input features, which can negatively affect the classification accuracy in many cases. Rough sets and scatter search are used for attribute reduction along with two classifiers, AELC and KMC. The performance of the proposed approach was evaluated using eleven different benchmarks. The classification of the used two classifiers was compared before and after attribute reduction. For completeness purposes, the classification results of other three classifiers were presented for the evaluation datasets without attribute reduction. The evaluation results supported the effectiveness of the proposed approach in removing redundancies of the input feature space.

7. REFERENCES

[1] S. Abdullah, L. Golafshan, and M. Z. Nazri. Re-heat simulated annealing algorithm for rough set attribute reduction. *International Journal of the Physical Sciences*, 8(6):2083–2089, 2011.

[2] C. Blake and C. Merz. University of california at irvine repository of machine learning databases, 1998. http://www.ics.uci.edu/mlearn/MLRepository.html.

[3] M. Bramer. *Principles of Data Mining*. Springer, Verlag London, 2007.

[4] I. D. Falco, A. D. Cioppa, and E. Tarantino. Facing classification problems with particle swarm optimization. *Applied Soft Computing*, 7(3):652–658, 2007.

[5] U. Fayyad and R. Uthurusamy. Evolving data into mining solutions for insights. *Communications of the ACM*, 45(8):28–31, 2002.

[6] S. García, A. Fernández, J. Luengo, and F. Herrera. A study of statistical techniques and performance measures for genetics-based machine learning: accuracy and interpretability. *Soft Computing*, 13(10):959–977, 2009.

[7] F. Glover. Surrogate constraints. *Operations Research*, 16:741–749, 1968.

[8] F. Glover. A template for scatter search and path relinking. *Artificial Evolution. In: J.K. Hao, E. Lutton, M. Schoenauer and D. Snyers, Lecture Notes in Computer Science 1363. Springer*, pages 125–137, 1998.

[9] F. Glover, M. Laguna, and R. Marti. *Scatter search*, pages 519–537. Advances in Evolutionary Computation: Theory and Applications, A. Ghosh and S. Tsutsui. (Eds.), Springer-Verlag, New York, 2003.

[10] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

[11] A.-R. Hedar, J. Wang, and M. Fukushima. Tabu search for attribute reduction in rough set theory. *Soft Computing*, 12(9):909–918, 2008.

[12] A. M. Ibrahim, A. Hedar, A. Abdel-Hakim, and A. Sewisy. *k*-means cloning: Adaptive spherical k-means clustering. submitted.

[13] A. M. Ibrahim, A. Hedar, A. Abdel-Hakim, and A. Sewisy. *k*-means mutant: Adaptive ellipsoidal k-means clustering. submitted.

[14] A. Jain and B. Chandrasekaran. Dimensionality and sample size considerations in pattern recognition in practice. *In: Krishnaiah, P.R., Kanal, L.N. (Eds.)*, 2:835–855, 1982.

[15] R. Jensen and Q. Shen. Semantics-preserving dimensionality reduction: Rough and

fuzzy-rough-based approaches. *Knowledge and Data Engineering, IEEE Transtion*, 16:1457–1471, 2004.

[16] D. Karaboga and C. Ozturk. A novel clustering approach: Artificial bee colony (abc) algorithm. *Applied Soft Computing*, 11:652–657, 2011.

[17] M. Laguna and R. Marti. *Scatter Search: Methodology and Implementations in C*. Kluwer Academic Publishers, Boston, 2003.

[18] T. Lin, Y. Yao, and L. Zadeh. *Data Mining, Rough Sets and Granular Computing*. Springer, Verlag, Berlin, 2002.

[19] A. Liu and Q. Zhu. Automatic modulation classification based on the combination of clustering and neural network. *The Journal of China Universities of Posts and Telecommunications*, 4(18):13–19, 2011.

[20] T. Mitchell. Machine learning and data mining. *Communications of the ACM*, 42(11):30–36, 1999.

[21] N. Nouaouria, M. Boukadoum, and R. Proulx. Particle swarm classification: A survey and positioning. *Pattern Recognition*, 46:2028–2044, 2013.

[22] S. Olafsson, X. Li, and S. Wu. Operations research and data mining. *European Journal of Operational Research*, 187(3):1429–1448, 2008.

[23] S. C. PANDEY. New facet of honey bees dancing language for mining the induction rules. *Sādhanā*, 43(3):35, 2018.

[24] Z. Pawlak. Rough sets. *International Journal of Computer and Information Science*, 11:341–356, 1982.

[25] Z. Pawlak. *Rough Sets: Theoretical Aspects of Reasoning about Data*. Kluwer Academic Publishers, Dordrecht, 1991.

[26] Z. Pawlak and A. Skowron. *Rough set methods and applications: New developments in knowledge discovery in information systems*. in Studies in Fuzziness and Soft Computing, Physica-Verlag, Berlin, 2000.

[27] J. Peter. A grafical aid to the interpertation and validation of cluster analysis. *Computational and Applied Mathematics*, 20:53–65, 1987.

[28] G. Peters, P. Lingras, D. lzak, and Y. Yao. *Rough Sets: Selected Methods and Applications in Management and Engineering*. Advanced Information and Knowledge Processing, Springer, Verlag London, 2012.

[29] M. T. S. K. Mandal, F. T.S. Chan. Leak detection of pipeline: An integrated approach of rough set theory and artificial bee colony trained svm. *Expert Systems with Applications*, 39:3071–3080, 2012.

[30] M. Shen, M. Peng, and H. Yuan. Rough set attribute reduction based on genetic algorithm. *Advances in Information Technology and Industry Applications*, 136:127–132, 2012.

[31] D. J. Sheskin. *Handbook of parametric and nonparametric statistical procedures*. crc Press, 2003.

[32] R. Swiniarski and A. Skowron. Rough set methods in feature selection and recognition. *Pattern Recognition Letters*, 24:833–849, 2003.

[33] R. W. Swiniarski and A. Skowron. Rough set methods in feature selection and recognition. *Pattern recognition letters*, 24(6):833–849, 2003.

[34] A. Unler and A. Murat. A discrete particle swarm optimization method for feature selection in binary classification problems. *European Journal of Operational Research*, 206:528–539, 2010.

[35] J. Wang, A. Hedar, S. Wang, and J. Ma. Rough set and scatter search metaheuristic based feature selection for credit scoring. *Expert Systems with Applications*, 39:6123–6128, 2012.

[36] J. Wang, Q. Zhang, A.-R. Hedar, and A.-M. M. Ibrahim. A rough set approach to feature selection based on scatter search metaheuristic. *Journal of Systems Science and Complexity*, 27(1):157–168, 2014.

[37] B. Xue, M. Zhang, W. N. Browne, and X. Yao. A survey on evolutionary computation approaches to feature selection. *IEEE Transactions on Evolutionary Computation*, 20(4):606–626, 2016.