Multi-Objective Scatter Search using Cuckoo Search with Modified Hierarchical Agglomerative Clustering for Resolving Clustering Problems

S. Karthikeyan[1] E. J. Thomson Fredrik[2]
[1]Research Student [2]Associate Professor
[1]Department of Computer Science [2]Department of Computer Applications
[1,2]Karpagam University, Coimbatore-21, India

Abstract— Data clustering is a vital concept of mining as it partitions the given dataset into meaningful set of clusters based on data similarity. This concept enhances the computation efficiency in the data analysis processes. In the preceding researches, Particle Swarm Artificial Bee Colony (PSABC) and Hybrid Artificial Bee Colony-Firefly Algorithm (HABC-FA) were developed for clustering to solve the clustering problem; however the imbalanced dataset problem still prevailed. This was overcome by the introduction of Multi-Objective Scatter Search Simulated Annealing with Hierarchical Agglomerative Clustering (MOSSSA-HAC) approach. But the feature selection in this approach has slow convergence speed. Hence in this paper, the Multi-Objective Scatter Search with Cuckoo Search algorithm with Modified Hierarchical Agglomerative Clustering (MOSSCS-MHAC) is proposed for resolving the convergence problem. In the initial stage, the Modified KNN is used for pre-processing followed by the under sampling process for error reduction. Then the multi-objective feature selection is performed using MOSSCS while the final clustering is achieved using MHAC algorithm. This approach provides optimal clustering with high accuracy. The experimental results show that the proposed MOSSCS-MHAC provides high values of precision, recall and f-measure than the existing algorithms.

Key words: Data Clustering, Particle Swarm Artificial Bee Colony, Hierarchical Agglomerative Clustering, Scatter Search, Cuckoo Search

I. INTRODUCTION

There is a growing requirement for the way to extract knowledge from the data [1]. Clustering is a descriptive task which partition the dataset based on the predefined similarity measure [2]. Clustering is the grouping of a particular set of objects based on their characteristics, aggregating them according to their similarities. Regarding to data mining, this methodology partitions the data implementing a specific join algorithm, most suitable for the desired information analysis. This clustering analysis allows an object not to be part of a cluster, or strictly belong to it, calling this type of grouping hard partitioning. In the other hand, soft partitioning states that every object belongs to a cluster in a determined degree. More specific divisions can be possible to create like objects belonging to multiple clusters, to force an object to participate in only one cluster or even construct hierarchical trees on group relationships.

Clustering techniques have been widely used in machine learning, pattern recognition, medical etc. Number of clustering algorithms has been proposed for different requirements and nature of the data [3]. Partition based clustering (k-Modes and its initialization methods) [4], hierarchical clustering [5] model-based clustering (EAST algorithm) [6], density-based clustering [7], graph-based clustering, and grid-based clustering are some basic clustering algorithms with their advantages and disadvantages. It is hard to discover the distance measure between two categorical data objects, greater the distance between the clusters more separated will be the clusters [8]. One of the well-known clustering for categorical data is k-Modes algorithm for large datasets. The traditional way to treat categorical data is binary but does not do justice to the large value difference such as for the very low and very high the difference is same. Multi-objective clustering has gathered.

In the preceding researches, the hybrid model of PSABC, HABC-FA and MOSSSA-HAC algorithms were developed for efficient clustering. The PSABC algorithm is a combination of Particle Swarm Algorithm (PSO) and Artificial Bee Colony (ABC) Algorithm used for data clustering on benchmark problems. However the time complexity in PSO reduces the performance. This has been resolved using the HABC-FA algorithm for clustering to solve the clustering problems. The FA incorporates the flashing behaviour of fireflies to achieve the optimal clustering solution using in the phases of the ABC. However it does not handle the imbalanced dataset and hence it provides lower performance results. To avoid the abovementioned issues, in the proposed system, the Multi-Objective Scatter Search Simulated Annealing with Hierarchical Agglomerative Clustering (MOSSSA-HAC) approach. The proposed system contains the phases such as pre-processing using Modified k-Nearest Neighbour (MKNN), under sampling based clustering, multi objective feature selection using scatter search with simulated annealing and HAC. The pre-processing step is used to increase the clustering accuracy by replacing the missing values. The under sampling is focused to avoid the error rates and obtained reduction dataset. The MOSSSA algorithm is used to produce optimal feature dataset. The HAC provides high quality clustering results rather than preceding research. However this approach has less convergence speed which can be improved by modifying the feature selection and clustering processes.

In this paper, the pre-processing is performed using MKNN as in MOSSSA-HAC while the feature selection is performed using Multi-Objective Scatter Search with Cuckoo Search algorithm. This approach enhances the optimal clustering performance. Then the final clustering operation is performed by Modified HAC algorithm. The remainder of this article is organized as: Section 2 describes the related research works. Section 3 explains the proposed research

methodology while the section 4 presents the evaluation results. Section 5 makes a conclusion of this research paper.

II. RELATED WORKS

For a data set with nontrivial size, it has been found difficult to identify the optimal clustering solution [9]. The problem is made even more difficult when the proper cluster number in the data set is unknown beforehand. Among various stochastic optimization techniques, a prominent approach is to use Genetic Algorithms (GAs) [10]. Maulik et al. [11], Tseng et al. [12] and Sarafis et al. [13] describe several methods based on traditional GAs for automatic data clustering.

GA and its variants, which are known to be effective for NP-hard problems, have therefore been widely employed to approach the clustering problem. Some of the early methods generally differ in the encoding scheme to represent the clustering solution and/or genetic operators (to generate new candidate solutions). In [14], Hruschka et al. devised a method attempting to improve the grouping genetic algorithm. In this method, an integer array is used to represent the clustering solution, such that the ith value in the array denotes the cluster membership of ith object. Both mutation and crossover operation in the method are redefined to cope with redundancy and context insensitivity issues of the encoding scheme.

In [15], Korkmaz designed a crossover operator called group-crossover to exploit the potential of using Linear Linkage Encoding (LLE) scheme for automatic data clustering. The LLE encoding scheme is also based on an integer array. However, the ith value, say j, in the array is interpreted as a link between data objects i and j, and they will be assigned to the same cluster in the resulting solution. In [16], Handl et al. adopted the LLE scheme with the uniform crossover and neighborhood-biased mutation to perform automatic data clustering by simultaneously optimizing two clustering criteria. While a real-value array, which is used to represent the cluster centers, was employed as the underlying encoding scheme to search for clusters. In this method, the crossover operation is devised to work by exchanging randomly selected sub-arrays. Comparing to the clustering methods mentioned above, the GA based clustering methods usually require more computational time. However, the reward of such an approach is that it can generally deliver more promising clustering solutions and makes no a priori assumption about the number of clusters.

Satapathy & Naik [17] proposed a new approach to using TLBO to cluster data. It is shown how TLBO can be used to find the centroids of a user specified number of clusters. The new TLBO algorithms are evaluated on some datasets and compared to the performance of K-means and PSO clustering. Karaboga & Ozturk [18] proposed the use of artificial Bee Colony (ABC) optimization algorithm, to classification benchmark problems.

Chen & Ye [19] proposed particle swarm optimization algorithm-based technique, called PSO-clustering, for cluster analysis to search the cluster center in the arbitrary data set automatically. PSO can search the best solution from the probability option of the Social-only model and Cognition-only model. This method is quite simple and valid, and it can avoid the minimum local value.

Sheng et al [20] proposed adaptive multi-sub-population competition (AMC) and multi-niche crowding and incorporated into a memetic algorithm to tackle the clustering problem. The AMC mechanism is developed to ensure a diverse search over solution subspaces corresponding to different numbers of clusters while allowing more promising subspaces to be more intensively searched. In this mechanism, the amount of individuals to be migrated between subpopulations is adaptively controlled according to the performance of subpopulations as well as the diversity of cluster numbers in population. Further, the migration is restricted to occur between subpopulations with relatively similar performances. Additionally, subpopulations with different performances are devised to search their corresponding subspaces with different exploration powers. This is achieved by dynamically adjusting parameter values of a multi-niche crowding method to form and maintain diverged niches of high fitness within the subpopulation. Though these techniques seem to be efficient, the convergence speed is a vital issue which is not satisfactory. The proposed research methodology in this article aims at overcoming this issue with significant performance improvement.

III. PROPOSED METHODOLOGY

In the proposed method, the Multi-objective Scatter search with Cuckoo search (MOSSCS) is employed along with Modified Hierarchical Agglomerative Clustering (MHAC) approach. The MOSSCS efficiently selects the best features while the MHAC clusters the data based on them.

The clustering problem has been generally evaluated based on the distance measures. From the numerical data attributes, the distance measures are computed using Euclidian, Manhattan, or maximum distance measure. However the detection of the measure for categorical attributes is difficult. Then the problem of identifying the number of clusters is performed without knowledge of the number of class labels. The analysis of the number of clusters is necessary to produce accurate results which would not be possible with the hierarchical approach as if a tuple gets a wrongly merged in a cluster, and then the process cannot be reversed. The partitional approach clusters the data and the selection of significantly best cluster after a few iterations. The under-sampling process is performed for the tackling of imbalanced dataset problem.

A. Pre-Processing

As in MOSSSA-HAC approach, the pre-processing of the input data is carried out using Modified k-Nearest Neighbor (MKNN) [21]. The main idea of the presented method is assigning the class label of the data according to K validated data points of the train set. In other hand, first, the validity of all data samples in the train set is computed. Then, a weighted KNN is performed on any test samples. In the MKNN algorithm, every sample in train set must be validated at the first step. The validity of each point is computed according to its neighbors. The validation process is performed for all train samples once. After assigning the validity of each train sample, it is used as more information about the points. Inorder to validate the points, the nearest neighbors are considered. Each of the K samples is given a weighted vote that is usually equal to some decreasing function of its

distance from the unknown sample. These weighted votes are then summed for each class, and the class with the largest total vote is chosen. This distance weighted KNN technique is very similar to the window technique for estimating density functions. Thus the MKNN performs significantly to enhance the data clustering.

B. Cluster based under-Sampling

Assume that the number of samples in the class-imbalanced dataset is N, which includes majority class samples (MA) and minority class samples (MI). The size of the dataset is the number of the samples in this dataset. The size of MA is represented as $Size_{MA}$, and $Size_{MI}$ is the number of samples in MI. In the class-imbalanced dataset, $Size_{MA}$ is far larger than $Size_{MI}$. For our under-sampling method SBC (under-sampling based on clustering), we first cluster all samples in the dataset into K clusters. In the experiments, we will study the performances for the under-sampling methods on different number of clusters [22].

Let the number of majority class samples and the number of minority class samples in the ith cluster ($1 \leq i \leq K$) be $Size^i_{MA}$ and $Size^i_{MI}$, respectively. Therefore, the ratio of the number of majority class samples to the number of minority class samples in the ith cluster is $Size^i_{MA}/Size^i_{MI}$. Suppose the ratio of $Size_{MA}$ to $Size_{MI}$ in the training dataset is set to be m:1 (m = 1). The number of selected majority class samples in the ith cluster is shown as

$$SSize^i_{MA} = (m \times Size_{MI}) \times \frac{Size^i_{MA}/Size^i_{MI}}{\sum_{i=1}^{K} Size^i_{MA}/Size^i_{MI}} \quad (1)$$

In the above expression, $m \times Size_{MI}$ is the total number of selected majority class samples that we suppose to have in the final training dataset. $\sum_{i=1}^{K} Size^i_{MA}/Size^i_{MI}$ is the total ratio of the number of majority class samples to the number of minority class samples in all clusters. This expression determines that more majority class samples would be selected in the cluster which behaves more like the majority class samples. In other words, $SSize^i_{MA}$ is larger while the ith cluster has more majority class samples and less minority class samples.

If there is no minority class samples in the ith cluster, then the number of minority class samples in the ith cluster (i.e., $Size^i_{MI}$) is regarded as one, that is, assumed that there is at least one minority class sample in a cluster. After determining the number of majority class samples which are selected in the ith cluster ($1 \leq i \leq K$), the majority class samples in the ith cluster are randomly chosen. The total number of selected majority class samples is about $m \times Size_{MI}$ after merging all the selected majority class samples in each cluster. Finally, the whole minority class samples are combined with the selected majority class samples to construct a new training dataset.

C. Multi Objective Feature selection using Scatter Search with Cuckoo Search algorithm

Multi-objective scatter search algorithm has been developed for combinatorial problems and has been enhanced using the Cuckoo search [23]. The improvement provides Scatter Search with random exploration for search space of problem and more of diversity and intensification for promising solutions. The proposed MOSSCS is applied for the selection of best features in the dataset. The algorithm must use data values on to find the best features to fulfill several different criteria.

The improvement to SS algorithm was accomplished by using nature inspired swarm intelligent algorithm, which is Cuckoo Search. Cuckoo search algorithm has proven its ability in solving some combinatorial problems and finding the nearest global optimum solution in reasonable time and good performance. Because the SS algorithm is composed of several steps, there will be several places to improve the SS algorithm. However, by the applied experiments, Subset Generation Method, Improvement Method and Reference Set Update Method are the most effective steps in improving the SS algorithm.

For improving the SS algorithm, the time is the big problem that is found in the Improvement Method. Where the Improvement Method is applying on all populations rather than to each new solution produced from Combination Method, so this will take a large amount of time, this will affect the SS algorithm as one of the meta-heuristic algorithms that the main goal of it in solving the problems is to find the optimal solution in reasonable time.

However, when trying to improve the SS algorithm in Reference Set Update Method in SS algorithm, the results were good and in reasonable time. The steps of CS will take its solutions from steps in SS, which is Reference Set Update Method and explore more of solutions and retrieve the best solutions reached to complete SS steps. In Reference Set Update Method, RefSet1 of b1 of the best solutions and RefSet2 of b2 of diversity of solutions will be chosen. RefSet1 will enter to the new steps that added from CS to SS. The new steps provide a more diversity to the RefSet1 which is benefit from the neighborhood search in the cuckoo search steps. Also the updated RefSet will contain more enhanced solutions than the old because the substitution operator forms the cuckoo solutions.

1) Algorithm 1: MOSSCS

- Input: $Psize, PRsize, NDsize, Ti, Tcr, Tstop, n_e,$;
- Output: Best Features

1) Initialize the population Pop using Diversification Generation Method.
2) Apply the Improvement Method to the population.
3) Reference Set Update Method (Good solutions for RefSet1 and Diversity solutions for RefSet2)
4) While (itr < MaxItr) do
5) While (Reference set is changed) do
6) Get a cuckoo randomly by Levy flight (from RefSet)
7) Evaluate its quality/fitness Fi
8) Choose a nest among n (say j) randomly
9) If (Fi > Fj) replace j by the new solution;
10) Fractions (pa) of worse nests are abandoned and new ones are built;
11) Keep the best solutions (or nests with quality solutions to substitute the RefSet);
12) Subset Generation Method
13) While (subset-counter < > 0) do
14) Solution Combination Method
15) Improvement Method
16) Reference Set Update Method;
17) End while
18) End while
19) End while

D. Modified Hierarchical Agglomerative Clustering

The idea of the Modified HAC [24] is to cluster the data using the item-based clustering approach based on the traditional HAC algorithm. Many existing clustering systems use term-based clustering approach. They construct data representation matrix based on the data frequency. According to data matrix, they choose the right similarity measures and then calculate similarity between clusters. The right choice of similarity measure and clustering algorithm is very important for clustering process. Moreover, constructing data matrix based on term-based or item-based approach is the main part of the clustering because this can affect the accuracy of clustering. Unlike existing systems, the proposed method constructs data matrix without considering counts of data By neglecting the data counts, the proposed method can merge the similar clusters into the same cluster efficiently than the existing systems with less processing time.

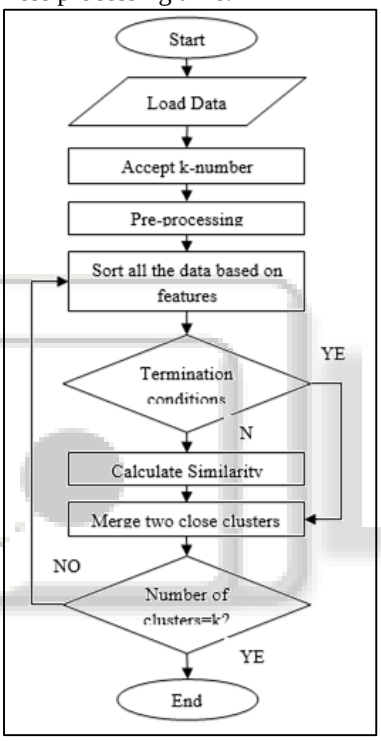

Fig. 1: Modified HAC flow diagram

The proposed method shown in Figure 1 mainly consists of three phases: data collection, data preprocessing and clustering process. In the proposed method, dataset which need to be clustered are loaded. After that, those data are pre-processed. After preprocessing data, the MHAC algorithm is performed. In MHAC algorithm, the number of cluster, k, is specified. After that data are sorted and the clusters are formed. If the data are similar, they are clustered together. The Jaccard coefficient is used to calculate similarity between the smallest cluster and all other clusters. Based on the Jaccard similarity scores, the two most similar clusters are merged into the same cluster. Thus efficient clustering is achieved.

1) Algorithm 2: Modified HAC

- Input: Data with features
- Output: Optimal clusters

Accept k-input number.
Sort all data based on the features
Clusters obtained
If $(X \cap Y) = X$ *or* $(X \cap Y) = Y$ then
Merge these two closet clusters
Else
Calculate similarity measure
Merge the most similar (closest) two clusters
Repeat until k-clusters
End

Thus the efficient clustering is achieved with combination of MOSSCS and MHAC algorithms. It is executed especially when the amount of dataset attributes in a given datasets is very large. As clustering results can characterize the basis for distribution of the whole data sets, clustering is helpful to aid supervised classification of datasets based on their selection of features. Thus, clusters can be used to select optimal and useful features also subsequently to add into sample datasets to improve the performance of classification in clustering process. Thus the classification steps used to increase the convergence speed of the clustering approach. The evaluation results in the following section provide the justification of this theory.

IV. Performance Evaluation

A. Data Sources

The proposed method has been evaluated through three data sets from different knowledge fields for the purposes of evaluating the performance and effectiveness of proposed MOSSCS-MHAC algorithm. The data sets are available from UCI Machine Learning Repository. Table 1 presents a description of the tested data sets, including the number of instances, number of features for each data set.

Bench Mark Datasets	No. of Samples	No. of features of the data object in the dataset
Fisher's iris dataset	50	Sepal length, sepal width, petals length and width
Thyroid dataset	175	Euthyroid, Hyperthyroidism, and Hypothyroidism Patients
Wisconsin breast cancer dataset	459	Clump Thickness, Cell Size Uniformity, Cell Shape Uniformity, Marginal Adhesion, and Mitoses.

Table 1: Sample of Benchmark Datasets

The Iris data set consists of 50 samples from each of three species of Iris. Four features were measured from each sample: the length and the width of the sepals and petals. The Thyroid dataset consist of 175 samples, together with three features such as euthyroid, hyperthyroidism, and hypothyroidism patients. The Wisconsin breast cancer dataset comprises of 459 samples, with six features such as clump thickness, cell size uniformity, cell shape uniformity, marginal adhesion, and mitoses. Based on the combination of datasets features, this proposed work developed an efficient clustering approach MOSSCS-MHAC for solving the clustering problem and the performance of this proposed algorithm compared with algorithm like MOSSSA-HAC, HABC-FA and PSABC.

B. Statistical criteria used in the clustering process

The statistical measures that follow are exploited for evaluating the results and compared generated results with MOSSCS-MHAC by various clustering algorithms.

1) Minimization of outline within criteria (OLW)

This criteria equation is based on data objective collective within groups' dimensional matrix W. The collective within data cluster object D- Dimensional matrix W is defined as,

$$\sum_{k=1}^{K} W_k \quad (2)$$

Where $W_k$refers to the variance matrix of the data object assigned to cluster C_k, where $k = \{1, \dots, K\}$.

$$W_k = \sum_{i=1}^{n_k} \left(\overrightarrow{O_l^k} - \overrightarrow{O^k}\right)\left(\overrightarrow{O_l^k} - \overrightarrow{O^k}\right)^T \quad (3)$$

where $\overrightarrow{O_l^k}$ indicates the i^{th} data object in cluster $C_k$and n_k refers to the number of objects in cluster C_k. and $\overrightarrow{O^k} = \frac{\sum_{i=1}^{n_k} \overrightarrow{O_l^k}}{n_k}$ indicates the vector of the centroid for the cluster C_k.Where K is no. of groups or clusters on the basis of certain similarity (distance) metric between the data points of the datasets. A set of N data objects has to get clustered in the process of clustering.

2) Maximization of variance ratio criteria

This criterion is dependent on data collective within group's D-dimensional matrix W and between the group D-dimensional matrixes*B*. The between dimensional matrix B is defined as per equation (2)

$$B = \sum_{k=1}^{K} n_k \left(\overrightarrow{O^k} - \overrightarrow{O}\right)\left(\overrightarrow{O^k} - \overrightarrow{O}\right)^T \quad (4)$$

Where,

$$\overrightarrow{O} = \frac{\left(\sum_{i=1}^{N} \overrightarrow{O_l}\right)}{N}$$

Therefore the variance of criteria is VAR defined as follows,

$$VAR = \frac{\left(\frac{(trace(B))}{(K-1)}\right)}{\left(\frac{(trace(W))}{(N-K)}\right)} \quad (5)$$

The measurement of the efficiency of the algorithms is done based on the criterion that follows:

Mean best fitness value of OLW, VAR as expressed in equations (3) and (5). Successive percentage (%) (i.e. percentage of the number of runs) that attain the best known data objective function value over the no. of simulations. The benchmark datasets are taken into consideration for evaluating the performance of the algorithms.

3) Mean best fitness value and variance

The mean best fitness values and variance of criteria of the clustering analysis corresponding to the benchmarked datasets are listed in the Table 2 and Table 3. The result indicates that the MOSSCS-MHAC provides best solution for clustering in both of the parameters such as mean best fitness value and VAR.

Datasets	Mean best Fitness value of OLW			
	PSABC	HABC-FA	MOSSSA-HAC	MOSSCS-MHAC
Fisher's iris dataset	69.25	69.75	69.67	70.02
Wisconsin breast cancer dataset	69.34	68.88	70.75	71.32
Thyroid dataset	68.58	68.65	70.89	72.11

Table 2: Mean best fitness value in the clustering analysis of benchmark datasets

Figure.2 illustrate that the performance i.e. mean best fitness value from the clustering processes.

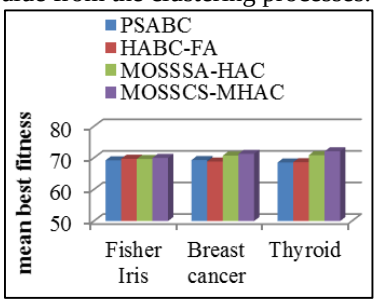

Fig. 2: Comparison of mean performance of the proposed and existing clustering processes through Datasets

Table 3 shows the VAR criterion of the clustering analysis of benchmark datasets fitness values of the algorithms using in the clustering analysis of the given datasets.

Datasets	Variance			
	PSABC	HABC-FA	MOSSSA-HAC	MOSSCS-MHAC
Fisher's iris dataset	42.95	44.79	46.73	47.32
Wisconsin breast cancer dataset	44.15	44.85	45.16	46.21
Thyroid dataset	44.21	44.91	45.54	46.43

Table 3: VAR criterion of the clustering analysis of benchmark datasets

The figure.3 illustrate that the performance i.e. variance of the algorithms using in the clustering analysis of the given benchmark datasets.

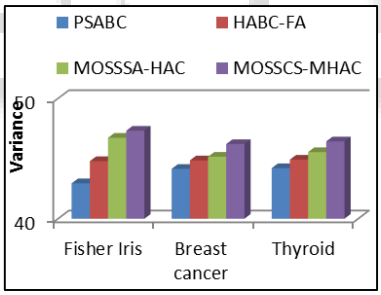

Fig. 3: Variance of the clustering Process comparing different algorithms

4) Success Percentage

The success rate which attain the reach best known objective function value in terms of percentage of number of runs (i.e., success %) is tabulated in Table.4 for benchmarked datasets. And Figure 4 illustrates the success rate of the newly introduced algorithm using in the datasets

Datasets	PSABC	HABC-FA	MOSSSA-HAC	MOSSCS-MHAC
Fisher's iris dataset	66	74	81	85
Wisconsin breast cancer dataset	90	92	94	96
Thyroid dataset	78	86	89	93

Table 4: Percentage of number of runs (i.e., success %) for benchmarked datasets

The figure 4 illustrates the Percentage of number of runs (i.e., success rate in %) for benchmark datasets using the proposed MOSSSA-HAC and existing methods.

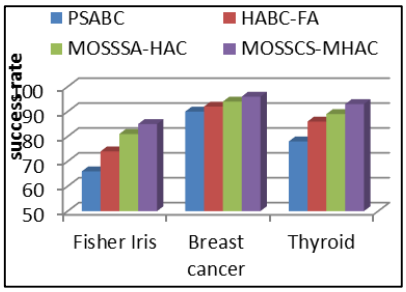

Fig. 4: Success rate of the Proposed Algorithm and the Existing Algorithm

C. Performance Analysis of proposed MOSSCS-MHAC

The evaluation of the proposed MOSSCS-MHAC clustering validation measures including recall, precision and F-measure. These three are generally used clustering validation measures for clustering process.

1) Recall

Recall value is obtained with the no. of most relevant feature dataset of cluster and the total no. of relevant feature datasets in cluster.

$$\text{Recall} = \frac{\text{No.of most relevant feature dataset of cluster}}{\text{Total No.of relevant feature datasets in cluster}} \quad (6)$$

Figure 5 shows the clustering performance comparison of the recall values of both proposed and existing methods for the given inputs datasets. It is found that the proposed MOSSCS-MHAC has higher values of recall due to the fact that the convergence is increased than the other methods.

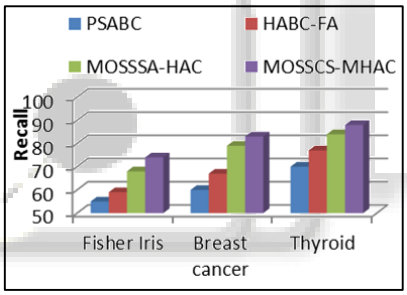

Fig. 5: Recall comparison

2) Precision

Precision value is computes with the total no. of relevant features in the datasets of cluster.

$$\text{Precision} = \frac{\text{No.of relevant dataset features clustered}}{\text{Total no.of dataset features clustered}} \quad (7)$$

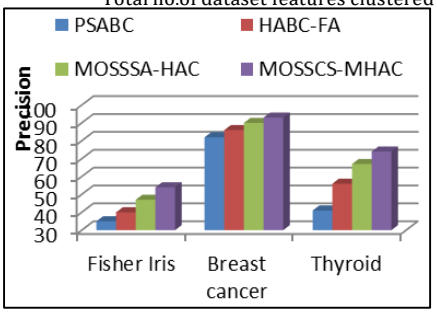

Fig. 6: Precision comparison

Figure 6 depicts the performance comparison of the precision values of both proposed and existing methods. The precision value is higher in the MOSSCS-MHAC algorithm than the other methods as the convergence and data imbalance problems are efficiently resolved.

3) F-Measure

The F-measure performance comparison for proposed and existing algorithms is calculated by combination of the precision and recall result values from the clustering process.

This work taken to treat each cluster results of a dataset, then calculate the recall and precision of that cluster for each given dataset and F-measure is calculated by using the following equation,

$$\text{F} - \text{measure} = \frac{2\ \text{Recall.Precision}}{\text{Precision+Recall}} \quad (8)$$

Figure 7 shows the f-measure comparison of the both proposed and existing methods. This result show that the MOSSCS-MHAC algorithm has higher values of f-measure than the other methods indicating the efficiency.

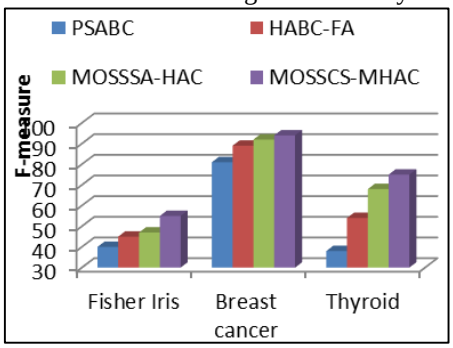

Fig. 7: F-measure comparison

From the observations, the Mean fitness value, VAR and success rate measure, all the four algorithms provide the best-known value to be within most no. of assessments and they have an increasing success rate to attain the optimal value for the dataset. The MOSSCS-MHAC outperformed the other methods. It is also observed that the convergence of the MOSSCS-MHAC is rapid for many of the benchmark datasets like iris, Wisconsin breast cancer and thyroid. The observation obtained from the OLW for VAR metric is that the performance of MOSSCS-MHAC is significant with regard to the benchmark datasets. Therefore it can be utilized for achieving efficient clustering results.

V. Conclusion

This paper presented an efficient clustering algorithm named as Multi-Objective Scatter Search with Cuckoo Search algorithm with Modified Hierarchical Agglomerative Clustering (MOSSCS-MHAC). This approach improves the selection of optimal features and also enhances the data clustering analysis process. The performance the proposed clustering algorithm has been evaluated over three benchmark datasets and the results are compared with that of the PSABC, HABC-FA and MOSSSA-HAC algorithms. The comparison results proved that the proposed clustering algorithm outperformed the other methods with high values of statistical measures namely mean, variance, success rate and performance measures namely precision, recall and f-measure. Though the efficiency of the proposed model has been justified, still there are some areas for improvement. The Hybrid feature selection will be applied with other efficient meta-heuristics and enhanced clustering concepts in the future.

REFERENCES

[1] Han, J, Kamber, M. (2001): Data Mining: Concepts and Techniques, Morgan Kaufmann Publisher, San Francisco, CA.
[2] Dongxia Chang, Yao Zhao, Changwen Zheng, Xianda Zhang, A genetic clustering algorithm using a message-based similarity measure Expert Systems with Applications, 39, (2012), 2194–2202.
[3] Sneha Antony, Jayarajan J N, T-GEN: A Tabu Search based Genetic Algorithm for the Automatic Playlist Generation Problem, Procedia Computer Science 46 (2015) 409 – 416
[4] Z. Huang, "A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets inData Mining", 1997.
[5] *Andreopoulos, A. An, X. Wang, "Hierarchical Density-Based Clustering of Categorical Data and a Simplification", PAKDD, 2007, pp. 11–22.*
[6] T. Chen, N.L. Zhang, Y. Wang, "Efficient model evaluation in the search-based approach to latent structure discovery", Proceedings of the Fourth European Workshop on Probabilistic Graphical Models (PGM-08), Vol. 8, 2008, pp. 57–64.
[7] Yinghua Lv, Tinghuai Ma , Meili Tang, Jie Cao, Yuan Tian , Abdullah Al-Dhelaan, Mznah Al-Rodhaan, An efficient and scalable density-based clustering algorithm for datasets with complex structures, Neurocomputing 171 (2016) 9–22.
[8] Arkajyoti Saha, Swagatam Das, Categorical fuzzy k-Modes clustering with automated feature weight learning, Neurocomputing 166 (2015) 422–435.
[9] Jain, A. K. (2010). Data clustering: 50 years beyond K-means. Pattern recognition letters, 31(8), 651-666.
[10] Deb, K., Pratap, A., Agarwal, S., & Meyarivan, T. A. M. T. (2002). A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE transactions on evolutionary computation, 6(2), 182-197.
[11] Maulik, U., & Bandyopadhyay, S. (2000). Genetic algorithm-based clustering technique. Pattern recognition, 33(9), 1455-1465.
[12] Tseng, V. S., & Kao, C. P. (2005). Efficiently mining gene expression data via a novel parameterless clustering method. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), 2(4), 355-365.
[13] Sarafis, I., Zalzala, A. M. S., & Trinder, P. W. (2002). A genetic rule-based data clustering toolkit. In Evolutionary Computation, 2002. CEC'02. Proceedings of the 2002 Congress on (Vol. 2, pp. 1238-1243). IEEE.
[14] Hruschka, E. R., Campello, R. J., & Freitas, A. A. (2009). A survey of evolutionary algorithms for clustering. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 39(2), 133-155.
[15] Korkmaz, E. E. (2006). A two-level clustering method using linear linkage encoding. In Parallel Problem Solving from Nature-PPSN IX (pp. 681-690). Springer Berlin Heidelberg.
[16] Handl, J., & Knowles, J. (2007). An evolutionary approach to multiobjective clustering. IEEE transactions on Evolutionary Computation, 11(1), 56-76.
[17] Satapathy, S. C., & Naik, A. (2011, December). Data clustering based on teaching-learning-based optimization. In International Conference on Swarm, Evolutionary, and Memetic Computing (pp. 148-156). Springer Berlin Heidelberg.
[18] Karaboga, D., & Ozturk, C. (2011). A novel clustering approach: Artificial Bee Colony (ABC) algorithm. Applied soft computing, 11(1), 652-657.
[19] Chen, C. Y., & Ye, F. (2012, May). Particle swarm optimization algorithm and its application to clustering analysis. In Electrical Power Distribution Networks (EPDC), 2012 Proceedings of 17th Conference on (pp. 789-794). IEEE.
[20] Sheng, W., Chen, S., Sheng, M., Xiao, G., Mao, J., & Zheng, Y. (2016). Adaptive Multisubpopulation Competition and Multiniche Crowding-Based Memetic Algorithm for Automatic Data Clustering. IEEE Transactions on Evolutionary Computation, 20(6), 838-858.
[21] Li, B., Yu, S., & Lu, Q. (2003). An improved k-nearest neighbor algorithm for text categorization. arXiv preprint cs/0306099.
[22] Yen, S. J., & Lee, Y. S. (2009). Cluster-based under-sampling approaches for imbalanced data distributions. Expert Systems with Applications, 36(3), 5718-5727.

