IJCTA, 9(36), 2016, pp. 503-515

Optimal Clustering Approach of Multi-Objective Scatter Search Simulated Annealing with Hierarchical Agglomerative Clustering for Clustering Problems

Karthikeyan S* and E.J. Thomson Fredrik**

Abstract : Data Mining is too possible to chunk away, concealed helpful acquaintance and data from profuse, imperfect, noisy, fuzzy and random realistic data. In data mining, the clustering method is one of the popular methods to be used. It is used to separate the data set into a significant set of reciprocally limited clusters with respect to relationship of data and it is used to create the more number of data in the same manner surrounded by a group and extra various among groups. In the existing work, in the clustering process the Particle Swarm Artificial Bee Colony (PSABC) and Hybrid Artificial Bee Colony-Firefly Algorithm (HABC-FA) are used to overcome the clustering trouble in the benchmark datasets like Fisher's iris data set. On the other hand, it does not manage the excessive data and therefore it gives the poor performance. To overcome the aforementioned troubles, in the present work, the Multi-Objective Scatter Search Simulated Annealing with Hierarchical Agglomerative Clustering (MOSSSA-HAC) method is introduced. In this present work, the process can be separated as phases like pre-processing using Modified k-Nearest Neighbour (MkNN), beneath sampling based clustering, multi objective feature selection using scatter search by way of replicated annealing and HAC. The first pre-processing process is used to enhance the accuracy of the clustering by changing the lost values. In the sampling stage it is observed to eliminate the error rates and acquired the lessening data set. The MOSSSA algorithm is used to generate optimal feature dataset. The HAC gives high quality clustering grades compared to the existing work. Finally, the proposed MOSSSA-HAC test result explains that the new MOSSSA-HAC method gives more accurate, recall and f-measure values contrast than preceding algorithms.

Keywords : Clustering, Artificial Bee Colony (ABC) algorithm, Modified k-Nearest Neighbour (MkNN), Multi-Objective Scatter Search Simulated Annealing (MOSSA) and Hierarchical Agglomerative Clustering (HAC).

1. INTRODUCTION

In the data analysis process more number of applications is used to verify how the data are prearranged. The steps of data analysis are observed on combining the data based on the correspondence measurements. The combination of a set of patterns into sub sets is called a cluster analysis or simply called clusters like the interpretation in the identical clusters are mostly unique to other cluster so they are to a pattern belonging to a various cluster. To represent the data there are more number of method can be used and examining the uniqueness among the data components and it generated a huge number of clustering approaches. This process is significant to the comparison among the supervised and unsupervised subjects. Both are unique in the sense that they categorize the data into groups based on similarity measurements [1].

* Research Scholar, Department of Computer Science Karpagam University, Coimbatore-21 s.karthics@gmail.com

** Associate Professor, Department of Computer Applications, Karpagam University, Coimbatore-21 thomson500@gmail.com

Clustering analysis process is the most significant function of the data mining and this analysis process is used as a stand-alone tool to acquire the information on the distribution of experiential features of the entire class, and it is observe on a particular class to do the additional analysis process [2]. Moreover, cluster analysis can be used as pre-processing steps of previous algorithms. Hence, cluster analysis process has develop into a data mining an extremely an energetic process of the research [3].

The number of require output clusters is assigned earlier than the algorithm describes the real cluster partition. The set of potential clustering partitions is very expensive for huge amount of data sets, therefore, metaheuristic method is participate a significant role in identifying the corresponding number of clusters as they iteratively try to enhance a person result with based to a present quality measure, hence it can search huge amount of spaces of candidate solutions. The evolutionary algorithm based on the metaheuristic method and it is used in optimization. This work is observed on assigning evolutionary computation methodology to cluster analysis trouble [4].

The trouble of cluster analysis has also been focused from multi-objective view using a various views of objectives with respect to acquire a set of non-dominated results as an alternative of a single solution. The researchers [5] introduce multi objective algorithms for the cluster analysis based on cluster ensembles. Non-ensemble multi-objective methodologies typically observed on conflicting similarity calculation like distance among the cluster centroids and calculate the dispersion within the clusters [6], [7], [8].

The *k*-means algorithm is generally used for the partition algorithms and it starts from an arbitrary first partition. It means the position of the various cluster centroids and holds the reassigning location of the centroids in the directions of the centre of collection of the cluster and it is represent in anticipation of a convergence criterion is met. For example, there are no more reassigning of cluster centroids. In the present work, the algorithm is introduced to create use of the aforementioned techniques to make the partitions of the data. The partitions size will be various and it is permitted by using variable length chromosomes in a same way [9]. A variable length chromosomes will be estimated using the popular multi objective evolutionary methodology NSGA – II [10] for which the characteristic formalisms.

The perception of multi-objective optimization method, while there is more number of objectives is concurrently optimized and there is no particular best possible solution. To a certain extent, there is set of best possible solutions, everyone assuming a certain trade-off between the objectives [11]. Similarly, a system is enhanced to overcome this type of troubles proceeds a set of best possible solutions, and can be gone to the user to select the best one that best solves particular troubles. On the other hand, that the user has the chance of selecting the solutions that represents the best trade off between the conflicting objective subsequent to investigate the more number of high quality solutions.

In this work, a Multi-Objective Scatter Search Algorithm with Simulated Annealing Algorithm (MOSSSA) is introduced to enhancing the multiple attributes on the clustering trouble. After that assign the Hierarchical Agglomerative Clustering (HAC) for efficient clustering is observed to cluster the most related data from the available datasets. In this present work, the MOSSSA-HAC methodology is used to enhance the accuracy, recall, f-measure values and precision for the optimal performance.

2. RELATED WORK

The Subtractive Clustering based Boundary Restricted Adaptive Particle Swarm Optimization method is suggested by Aher et al [12] for the clustering multi dimensional data. The Particle Swarm Optimization (PSO) is based on the swarm intelligence paradigms that have gained the widespread concentration in research [12]. The PSO algorithm is known as evolutionary computation method and it simulates the motion of flock of birds that performs a global search within a solution space. This algorithm creates good grades in difficult and multi peak trouble with a small number of attributes to regulate the speed and also the exact computation results and it is guide to be an optimization method in swarm intelligent area [13].

The clustering method is well-known method presented by Chişet al [14] that makes the effort to divide the data into individual groups that the data points and the characteristic is similar in order to a transfer point while the data points of various groups varied according to its characteristics. Every defined group are known as clusters. Therefore, the clusters are compacted of more number of unique data or objects in order to a transfer point. Cluster is mainly used by the engineering and science field as well as data compression, statistical data analysis, pattern recognition, data mining, artificial intelligence etc.,. a few applications like fingerprint recognition, document classification, handwritten character recognition and speech/speaker recognition need the use of clustering method with respect to decrease the training data amount or to identify the representative data points.

The Fast Balanced k-means (FBK) algorithm is presented by Sewisyet al [15]. The aim of clustering process is to collect a set of objects into clusters for the reason that the objects in the similar cluster are extremely unique other than the different objects in the other cluster. To undertake this trouble, the different kind of clustering algorithms have been implemented. Between them, for the large sale spherical data sets the efficient k-means clustering algorithm is used like information retrieval. To solve the insufficient of using the FBK algorithm the efficient clustering algorithm is introduced. To complete the process it is only needed a less computational time. This algorithm is also worked in the balanced data too.

The genetic algorithm is discussed by the Min et al [16] and it is a well known evolutionary algorithm, created by simulating the rule of survival of the fittest in natural surroundings. It also has the fitness measurements, genes coding, creating the initial population and estimate the evolutionary operation and so on. And it is also consists of crossover, mutation and selection. The Bee algorithm is introduced by Pham et al [17] and it is the optimization algorithm. It is motivated by the useful behaviour of honey bees to identify an optimal solution.

A hybrid methodology is introduced by ZeinEldin et al [18] and it is based on Simulated Annealing (SA) and Scatter Search (SS) to overcome the multi objectives optimization troubles. There are different types of experimental troubles are used to differentiate the performance of this method with other methodologies. The results demonstrate that the present method is effective and competitive with the other implemented methodologies.

3. PROPOSED METHODOLOGIES

The Multi-Objective Scatter Search with Simulated Annealing (MOSSSA) with Hierarchical Agglomerative Clustering (HAC) methodology is introduced in this proposed methodology.

3.1. Cluster problems

The identification of distance measure : For the mathematical parameters, distance calculations that can be used are standard equations such as Euclidian, Manhattan and highest distance measure. The entire three are special cases of Minkowski distance. Other than the classification of measure for definite parameters is complicated.

The number of clusters : In earlier days the number of cluster identification is a complex job because the number of class labels is not recognized. A suspicious analysis of number of clusters is essential to create accurate results. Moreover, it is identified that heterogeneous tuples may combined or same kind of tuples may be divided into more numbers. This could be catastrophic stipulation the methodology used is hierarchical. For the reason that, in hierarchical methodology if a tuples are combined in incorrectly in a cluster that action cannot be unfinished.

Despite the fact that, there is no possible way to calculate the number of clusters, there are some statistics that can be analysed to assist in the performance [19]. The required class labels are: for the original datasets, the data sharing has to be completed to understand where the class labels are?

Structure of database : The original data is not always holds the identifiable clusters. And also the order in which the tuples are arranged may affect the outcome while an algorithm is executed if the distance calculation used is not ideal. With a structure less data still the identification of corresponding number of clusters will not yield better outcome. For instance, the missing values can already present for the variables, tuples and thirdly, arbitrary in parameters and tuples. If a record has the entire values are lost, this is eliminated from the dataset. Suppose, a parameter has missing values in the entire tuples then that parameter has to be eliminated process is explained. To handle the missing values in the proposed system based on the mean and mode methodology by using the three cluster-based algorithms [20].

Types of attributes in a database : The databases may not require holding the distinctively mathematical or categorical parameters. They may also hold the other types like nominal, ordinal, binary and so on. So these parameters have to be converted to unconditional type to create measurements in easy.

Choosing the initial clusters : For partitioning methodology, it identifies the most important methodologies like k primary clusters to be arbitrary selected. A suspicious and complete learn of data is necessary for the identical. Moreover, the initial clusters are not appropriately selected, after that a small number of iterations it is identified that clusters may still be gone unfilled. Even if, in this work [21] discusses a furthest heuristic based methodology for measurements of centres.

Imbalanced dataset : The more number of class or the minimum number of class are available in the data set and therefore, the accuracy of the data set is decreased importantly. It has to be overtaken by cluster based sampling methodologies in this work.

3.2. Pre-processing

In this work, the pre-processing is used to eliminate the noise data from the particular data set. The altered K-Nearest Neighbour (MkNN) based algorithms choose the parameter with comparable to the parameter of interest to assign missing values. Assume the data A has one missing value in 1, this technique would identify the K other data or parameters which have a value present in experiment 1, with expression most similar to A in experiments 2-N. the total number of experiments is represented as N. a weighted average of values in experiment 1 from the K closest genes is then used as an calculation for the missing value in gene A.

Algorithm

```
1. Consider the input as X = $x1, x2….. xn$
2. Fix the threshold value T
3. Compute each input data X
4. Compare each data X with threshold T
5. If the data belongs to threshold T then consider into account G
6. Compute similarity and nearest points
7. Find out the missing values and fill the gaps
8. Else remove the points
```

3.3. Cluster based under-sampling

Down-sizing the popular class results in a defeat of data that possibly will result in excessively common regulations [22]. So to solve these demerits of the sampling methodologies Yen and Lee (2009) introduced cluster based beneath sampling. This methodology is to primary cluster of the entire training samples into K clusters (they have run the experiment with various K values to spectator the result) after that select the corresponding training samples from the copied clusters. The major thought is that there are various clusters in a dataset, and every cluster seems to have distinct features. Suppose a cluster has more number of class samples and minimum class samples means it will act like a majority class sample. However, if a cluster has additional alternative class samples and smaller amount preponderance class samples. The features are not available in these clusters of the majority class samples and act like a minority class samples. Hence, their methodologies choose an appropriate number of majority class samples from every cluster by assuming the ratio of the number of majority class samples to the number of minority class samples in the derived cluster.

The primary cluster entire data to K clusters. An applicable number (M) of majority class samples from every cluster are then chosen by the appropriate the ratio of the number of majority class samples to the number of minority class samples in the cluster. The number M is estimated by formula 1, and they arbitrary select the M numbers of majority class samples from every cluster. In the *i*th cluster ($1 \leq i \geq K$) the $\mathrm{Size}^{i}{}_{\mathrm{MA}}$ can written as follows:

$$\mathrm{Size}^{i}{}_{\mathrm{MA}} = (mx\,\mathrm{Size}_{\mathrm{MI}})x\frac{\mathrm{Size}^{i}{}_{\mathrm{MA}\,\mathrm{Size}^{i}_{\mathrm{MA}}}}{\mathrm{K}_{-1}\mathrm{Size}^{i}{}_{\mathrm{MA}\,\mathrm{Size}^{i}_{\mathrm{MA}}}} \tag{1}$$

This method may be appropriate for data sets where the class labels are confidently described and accurately reproduce the belongings of the labelled class. Other than, in some other cases, particularly for medical datasets, there is no assurance that the class labels are accurately reflecting the original features of that record.

This method to under-sampling is various to the methodology [23]. The entire majority class samples are available in one subset and the other subset has all the minority class samples. After that, the cluster majority class samples to K clusters (K>1) then create K subsets of majority class samples, while every clusters is measured to be one subset of the majority class. It is required to decrease the break among the numbers of majority class samples to the numbers of minority class samples. The entire majority sub sets are grouped individually with the minority class samples to create K various training data sets.

3.4. Multi Objective Feature selection using Scatter Search with Simulated annealing algorithm

Multi-objective optimization trouble holds more number of objectives that required to be reached at the same time. These kinds of troubles occur in more number of applications, sometimes in the challenging or incommensurable objectives functions have to be minimized simultaneously. This collection of optimization troubles has a various viewpoints differentiated to a single objective troubles. In this single objective optimization trouble there is only one global optimum, other than in multi-objective optimization having a set of solutions known as Pareto-optimal set, which are considered global optimum solutions [24]. Multi-objective optimization needed more computational effort compared to the single-objective optimization. Expect that the preferences are not related or totally unstated. The more number of single objective trouble solutions may be required to acquire a suitable final solutions.

The evolutionary methodology is also known as scatter Search (SS) is presented as a meta-heuristic for integer programming. It is based on the diversifying the search all the way through the solution space. This method can be execute a set of solutions named reference set (PR) and created by better and sparse solutions of the important population (P). These kinds of results are periodically grouped with the goal of creating novel solutions with good fitness; at the same time it managing the diversity. Moreover, an enhancement phase using local search is assigned.

- A set of initial solutions creates a diversification generation method.
- An enhancement approach that transfer a sample solution into the other more number of sample solutions.
- A reference set is used to update the approach and it manages a segment of the best solutions of the main set.
- A subset generation method that executes in the reference set in such a way as to choose the additional solutions with respect to it is grouped together.
- The grouping approach solution that transfer the solutions created by the subset creation approach into more number of grouping solution vector.

The SS method is a suitable methodology, because every module of its structure can be developed in various ways and to various levels of difficulty. For example, it is assumed that SS employs in good with population sizes (Psize) concerning characteristically at least 10 times the size of the reference set (PRsize) [25]. The reference set must be creating in specific attitude in intellect various stages. Some part of the solutions created the good solutions of the major population and the additional solution is the important distant solutions from the preceding ones. At last, in the designing aspect like the enhancement method can be developed using a huge amount of methodologies.

Another name of multi objective hybrid optimized algorithm is known as Multi-Objective Scatter Search with Simulated Annealing (MOSSSA), and it consists of extending the standard SS algorithm in three important aspects. Most of the objectives functions using Pareto-dominance idea, as in the MOMHs defined above. The Simulated Annealing as enhancement approach and it is considered as the second fact. The simulated annealing (SA) is a programming approach that creates to simulate the process by miss locating the atoms in a metal while it's fiery and after that it bit by bit chilled. The SA approach is used to optimize a solution by revealing it to high initial temperature (T_i), chilled it by way of a cooling rate (T_{cr}) in anticipation of the temperatire falls below a given threshold and it is represented as Tstop. There is a arithmetical demonstration about the convergence of SA approach is presence to the global optimum of a trouble while the temperature is reduced in bit by bit.

Algorithm

```
Input : Psize, PRsize, NDsize,Ti, Tcr, Tstop, $n_e$,;
P ← ϕ; PR ← ϕ; ND ← ϕ; eval_counter ← 0;
For ($P_{count}$ = 1 to $P_{size}$)
    P [$P_{count}$] ← Diversification_Method ();
    ND ← FAS (ND, P[$P_{count}$]);
    Improvement_Method (P[$P_{count}$]);
Sort the solutions of P according the objective function;
For ($PR_{count}$ = 1 to $PR_{size}$/2)
    PR [$PR_{count}$] P [$P_{count}$];
Sort the solutions of P according their distance to the distance to the solutions included in PR;
For ($PR_{count}$ = ($PR_{size}$/2) + 1 to $PR_{size}$)
    $PR[PR_{count}] \leftarrow P[PR_{count} - (\frac{PR_{size}}{2})]$;
Do
    New_solutions ← False
    NewSubsets ← Sub_Set_Generation_Method ();
    Do
        Select the next subset Z in Newsubsets;
        Trial ← Solution_CombinationMethod (Z);
        Trial2 ← trial
        ND ← FAS (ND, trial2)
        Improvement_Method(trail2);
        If (trial2<trial) then
            Trial ← trial 2;
        New_solutions ← true
        Delete Z from NewSubSets;
        For ($PR_{count}$ = 1 to $PR_{size}$)
            If (trial<PR[$PR_{count}$]) then
                PR [$PR_{count}$] then
                PR [PRcount]trial;
    While (NewSubSets ≠ ∅);
While (new_solutions = true) and (eval_counter < $n_e$));
```

3.5. Hierarchical Agglomerative Clustering

In this section, the agglomerative clustering algorithm is used to enhance the clustering accuracy by using numerous data. In the N cluster the Agglomerative clustering process is started and every one includes accurately one data point. A sequence of combination operations then followed, that ultimately forces the entire objects into the similar group. The Agglomerative algorithms start with the entire component as an individual cluster and combined into the sequentially bigger clusters. The identification of the similarity of pairs by using the following equation.

$$\mathrm{sim}(c_i, c_j) = \max_{x \in c_i ,} \mathrm{sim}(x, y) \tag{2}$$

The entire agglomeration clusters are created at a better distance among the clusters compared to the existing agglomeration and the one agglomeration choose to terminate the clustering process wither while the clusters are to take the long distance to be combined or a few number of clusters are available.

In the entire data, the algorithms is started the execution as a singleton cluster. At every step, the important same information corresponding to the similarity in the two clusters is combined into single clusters, generating one less cluster at the future level. In the seed set the iterative nearest neighbor classification (NN) algorithm is processed. The adjacent node will execute at every step of the set is included in the pattern, in anticipation of the adjacent node is no longer adequately same. There are three various types of methods are used in HAC and NN methodology for cluster-cluster or cluster-crime similarity namely Single Linkage (SL), Complete Linkage (CL) and Group Average (GA). The single linkage is an important pair of the data. The different pair of data is known as complete linkage. Group Average is used to the averaged pair wise similarity. Therefore, it is used to identify the exact data more efficiently rather than the existing algorithm for a given larger dataset.

HAC is one of the bottom-up clustering approaches while the clusters have sub clusters, which in turn have sub clusters and so on. The sample data reveal the hierarchical quality. In the entire single objects starts an agglomerative hierarchical clustering starts with each single object (sample) in a single cluster. After that, in every successive iteration, it agglomerates combines the nearest pair of clusters by gratifying some similarity criteria, in anticipation of the entire data is in one cluster.

Advantages : This method generates a correct objects and it may be useful data in the display function. A few amounts of clusters are created, which may be useful for discovery to calculate the uniqueness among the prototypes and data points, and it performed it good manner.

The grouping of clustering is implemented by the proposed methodology in this clustering process with the use of classification with respect to draw the best clustering result. A better subset of characteristic can not only enhance the accuracy of classification, and also it decreases the time to calculation conditions. It is processes particularly while the number of dataset parameters in a given dataset is very huge. In the entire set of data the clustering process is distributed and this process is useful to assist the classification of the datasets based on their appropriate characteristics. Therefore, the clusters can be used to choose the exact and helpful characteristic also consequently to sum into the sample datasets to enhance the performance of the classification process. Therefore, this step is used to improve the convergence speed of the clustering process.

4. EXPERIMENTAL RESULT

4.1. Data Sources

The present approach has been estimated all the way through the three data sets from various areas for the reason of estimating the performance and efficiency of new Hybrid ABC-FA algorithm. The UCI Machine Learning Repository containing the data sets in the cluster. The tested data sets, including the number of instance and the number of features for every data set details are explained in table 1.

Table 1

The sample of Bench mark Datasets

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.

The 50 samples are available in the Iris data set from the every of three species of Iris. Four parameters were calculated from every sample namely length, width, sepals and petals. The 175 samples are available in thyroid dataset, by combining these three characteristics like hyperthyroidism, hypothyroidism patients and euthyroid. There are 459 samples are comprised in the Wisconsin breast cancer dataset with six characteristic like cell size uniformity, marginal adhesion, cell shape uniformity, mitoses and clump thickness. This present work implemented an effective clustering methodology named as HABC-FA based on the datasets characteristics for resolve the clustering trouble and the performance of this present algorithm is distinguished with algorithm such as ABC and PSABC.

4.2. Statistical criteria used in the clustering process

The statistical calculations that follow are demoralized for estimating the results and compared generated results with HABC-FA by different clustering algorithms [26].

Minimization of outline within criteria (OLW)

This formula is based on the objective of data collection within the combination of dimensional matrix W. the grouping of the data cluster object D-Dimensional matrix W is described as follows:

$$\sum_{k=1}^{K} W_k \tag{3}$$

The variance matrix of the data object is applied to cluster C_k is represented as W_k, wherever k = {1, ... K}.

$$W_k = \sum_{i=1}^{n_k} (\overrightarrow{O_j^k} - \overrightarrow{O^k})\,(\overrightarrow{O_l^k} - \overrightarrow{O^k})^T$$

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 wherever K is number of groups or clusters on the foundation of convinced resemblance (distance) metric among the data points of the datasets. A set of N data objects has to get clustered in the process of clustering.

Maximization of variance ratio criteria

This Maximization of variance ratio is dependent in data cooperative in group's D-dimensional matrix W and among the collection D-dimensional matrixes. The among dimensional matrix B is definite as per equation (3)

$$B = \sum_{k=1}^{K} n_k (\overrightarrow{O^k} - \overrightarrow{O})\,(\overrightarrow{O^k} - \overrightarrow{O})^T \tag{4}$$

Where,

$$\overrightarrow{O} = \frac{\left(\sum_{i=1}^{N} \overrightarrow{O}_1\right)}{N} \tag{5}$$

Consequently the variance of criteria is VAR definite as follows,

$$\text{VAR} = \frac{\left(\dfrac{\text{trace(B)}}{(\text{K}-1)}\right)}{\left(\dfrac{\text{trace(W)}}{(\text{N}-\text{K})}\right)} \tag{6}$$

Based on the criterion the calculations of the efficiency of algorithm are achieved as follows:

The OLW, VAR mean best fitness value is represented in equations (4) and (6). The good percentage is obtained from the data objective function value in the excess of the number of simulation. The benchmark datasets are occupied into deliberation for estimating the performance of the algorithms.

Mean best fitness value and variance

In table 2 and 3 explains that the mean best fitness values and the variance of the clustering analysis appropriate to the benchmarked data sets. The outcome of the clustering process is indicates that the HABC-FA gives the good solution for clustering in both of the attributes like mean best fitness value and VAR.

Table 2

The Mean best fitness value in the clustering analysis of benchmark datasets such as Fisher's iris, Thyroid and Wisconsin breast cancer

Datasets	*PSABC*	*HABC-FA*	*MOSSSA-HAC*
	Mean best Fitness value of OLW	*Mean best Fitness value of OLW*	*Mean best Fitness value of OLW*
Fisher's iris dataset	69.25	69.75	69.67
Wisconsin breast cancer dataset	69.34	68.88	70.75
Thyroid dataset	68.58	68.65	70.89
Mean Best fitness Value of OLW	68.97	69.22	70.44

The mean best fitness value from the clustering processes is demonstrated in figure 1.

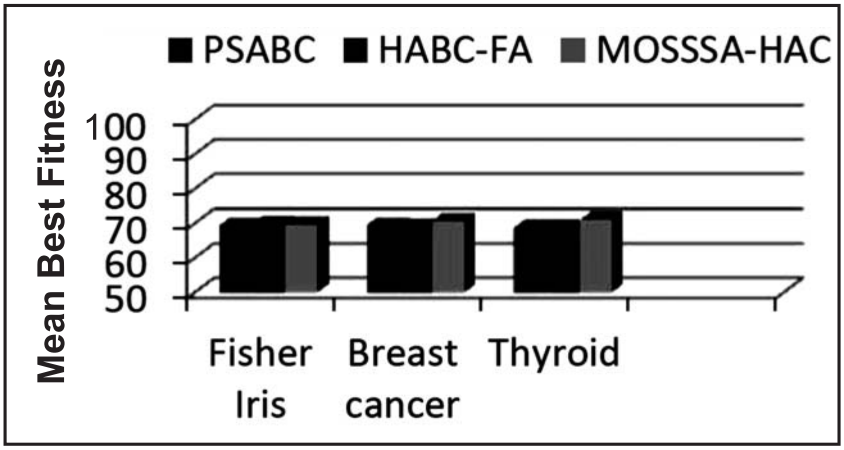

Figure 1: The comparison result of the mean performance of the proposed and existing clustering processes through Datasets

The clustering analysis benchmark process of the datasets fitness values of the algorithms is executed in VAR methods and it is explained in table 3.

Table 3

VAR criterion of the clustering analysis of benchmark datasets

Datasets	*PSABC*	*HABC-FA*	*MOSSSA-HAC*
	VAR	*VAR*	*VAR*
Fisher's iris dataset	42.95	44.79	46.73
Wisconsin breast cancer dataset	44.15	44.85	45.16
Thyroid dataset	44.21	44.91	45.54

The variance of the algorithms in the clustering analysis process by using the benchmark data sets is illustrated in figure 2.

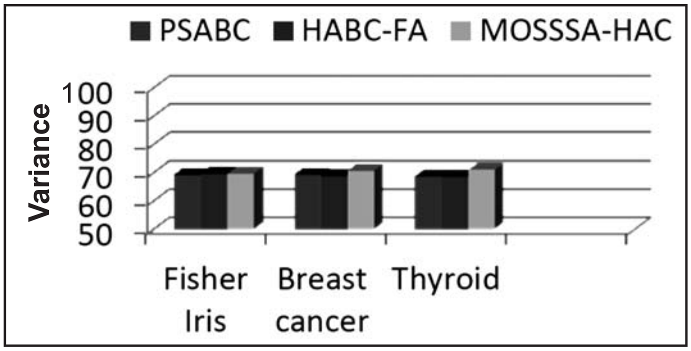

Figure 2: Variance of the clustering Process comparing different algorithms Success percentage

The accurate rate reach the best result is called as objective function value with respect to the percentage of number of process and this is explained in table 4. And the accurate rate of the new algorithm is illustrated in figure 3.

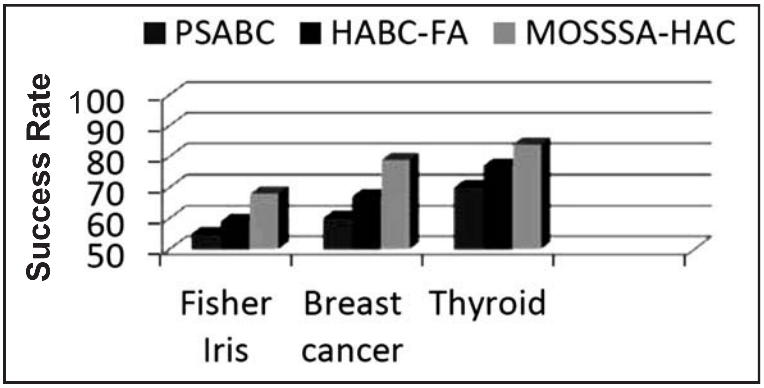

Figure 3: The success rate of the Proposed Algorithm and the Existing Algorithm

The process of MOSSSA-HAC and the preceding approach is illustrated in figure 3.

Table 4

Percentage of number of runs (*i.e.*, success %) for benchmarked datasets

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

4.3. Performance Analysis of proposed MOSSSA-HAC

The estimation of the present MOSSSA-HAC clustering justification calculations including the recall, precision and F-measure functions. These three types are commonly used in the clustering validation calculations.

Recall (RC)

The recall value is acquired with the more number of related characteristic dataset of cluster and the entire number of related feature datasets in cluster.

$$\text{Recall (RC)} = \frac{\text{No.of most relevant feature dataset of cluster}}{\text{Total No.of relevant feature datasets in cluster}} \tag{7}$$

The differentiation of the recall values of the clustering performance of both the present and preceding methods for the given input datasets are illustrated in figure 4.

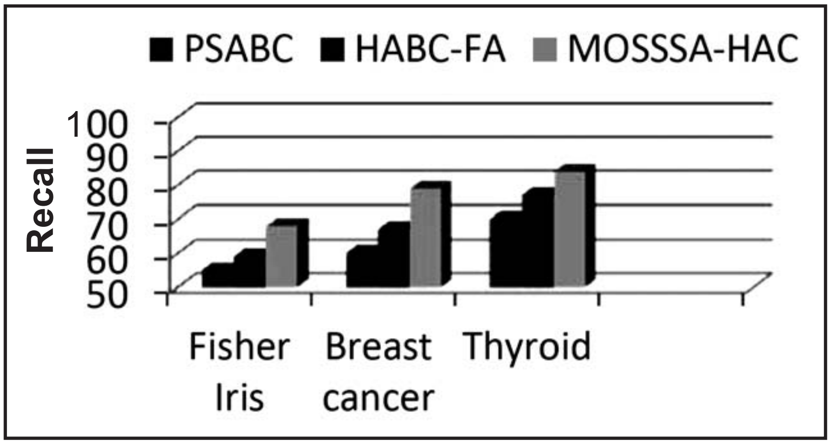

Figure 4: Recall comparison

Precision (PC)

The precision value (PC) is computes with the total no. of relevant features in the datasets of cluster.

$$\text{PC} = \frac{\text{No. of relevant feature dataset of clustered}}{\text{Total No. of dataset feature clustered}} \tag{8}$$

The differentiation of the precision values of the clustering performance of both the present and preceding methods for the given input datasets are illustrated in figure 5.

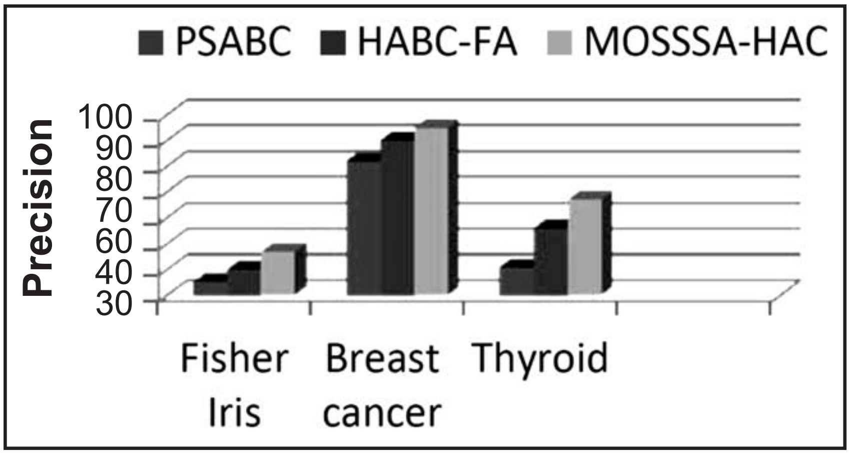

Figure 5: Precision comparison

F-measure

The F-measure performance differentiation for the present and preceding algorithms is measured by the grouping of the precision and recall outcome values from the clustering process.

The entire clusters are considered in this work, after that estimate the recall and precision of that cluster for all the specified dataset and F-measure is premeditated by using the following formula,

$$\mathrm{F - measure} \;=\; \frac{2\ \mathrm{RC, PC}}{\mathrm{PC + RC}}$$

The comparison of the f-measure of both the present and preceding methods is illustrated in figure 6.

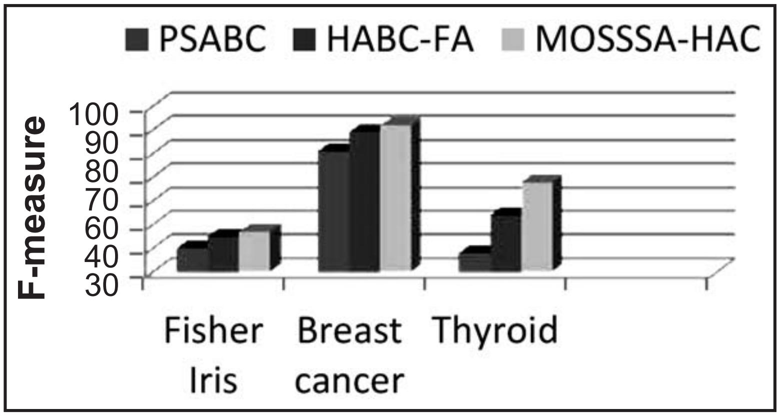

Figure 6: F-measure comparison

From the clarification, the three algorithms namely the Mean fitness value, VAR and success rate measure give the best known value to be within most number of assessments and they have an increasing success rate to obtain the optimal value for the dataset. The MOSSSA-HAC performance is compared with the other performance. It is also focused that the convergence of the MOSSSA-HAC is quick for more number of benchmark datasets such as iris, Wisconsin breast cancer and thyroid. The surveillance acquired from the OLW for VAR advantages is that the performance of MOSSSA-HAC is important with respect to the benchmark datasets. The MOSSSA-HAC algorithm has a higher possibility for identifying the significant optimal value while differentiated to PSABC, HAC-FA and MOSSSA-HAC. For that reason, the convergence of the MOSSSA-HAC for the Variance measure has a better speed while differentiated to other algorithms for more number of benchmark troubles. The present MOSSSA-HAC clustering algorithm performance is also calculated using the clustering process estimation attributes such as recall, precision and F-measure.

5. CONCLUSION

In this work, the MOSSSA-HAC algorithm is introduced to give the exact clustering process outcome. It integrates the basic characteristics of a Simulated Annealing (SA) algorithm grouped with PSABC, and it is improve the result for the clustering trouble. The comparison process is made by using the recall, precision and f-measures with the preceding algorithms using benchmark datasets. The present MOSSSA-HAC algorithm is generated high quality clustering process for the datasets. It is also focused the convergence of the MOSSSA-HAC is fast for more number of benchmark datasets such as iris, cancer and thyroid. The examination from the OLW for VAR features is that the performance of MOSSSA-HAC is extraordinary in order to the benchmark datasets. The MOSSSA-HAC algorithm has higher possibility to identify the needed accurate value in differentiation to preceding methods PSABC and HABC-FA. Therefore, the convergence of the present MOSSSA-HAC for the variance measure is quicker as compared to the other algorithms. Finally, the results shows that the present MOSSA-HAC algorithm performance is more efficient compared to the other one. In future work, the scalable clustering process can be worked and it can be assigned with hybrid clustering methodology to hold the outlier data more efficiently.

6. REFERENCES

1. Recio, Gustavo, and Kalyanmoy Deb. “Solving clustering problems using bi-objective evolutionary optimisation and knee finding algorithms.” *2013 IEEE Congress on Evolutionary Computation*. IEEE, 2013.
2. Maharaj E. A. Cluster of Time Series[J].Journal of Classifieation，2000，17(2):297-314
3. Guedalia I.D.，London M. ，Werman M. An on-line agglomerative clustering method for non-stationary data[J Neural Computation，1999，11(2).
4. A. K. Jain, M. N. Murty, and P. J. Flynn, “Data clustering: A review,” ACM Computing Surveys, vol. 31, no. 3, pp. 264–323, 1999.
5. K. Faceli, A. de Carvalho, and M. de Souto, “Multi-objective clustering ensemble,” in Hybrid Intelligent Systems, 2006. HIS ’06. Sixth International Conference on, 2006, pp. 51 –51.
6. J. Handl and J. Knowles, “Evolutionary multi-objective clustering,” Lecture notes in computer science, vol. 3242, pp. 1081–1091, 2004.
7. “Multi-objective clusterng detection with automatic determination of the number of clusters,” Technical report No. TR-COMPSYSBIO- 2004-02, UMIST, Department of Chemistry, Tech. Rep., 2004.
8. “An evolutionary approach to multiobjective clustering,” Evolutionary Computation, IEEE Transactions on, vol. 11, no. 1, pp. 56 –76, 2007.
9. K. Ripon, C.-H. Tsang, S. Kwong, and M.-K. Ip, “Multi-objective evolutionary clustering using variable-length real jumping genes genetic algorithm,” in Pattern Recognition, 2006. ICPR 2006. 18th International Conference on, vol. 1, 2006, pp. 1200 –1203.
10. K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, “A fast and elitist multiobjective genetic algorithm: NSGA-II,” Evolutionary Computation, IEEE Transactions on, vol. 6, no. 2, pp. 182 –197, 2002
11. CoelloCoello, C.A., Van Veldhuizen, D.A., Lamont, G.B.: Evolutionary Algorithms for Solving Multi-Ob je tive Problems. Kluwer A ademi Publishers, New York (2002).
12. SatyobrotoTalukder, ―Mathematical Modelling and Applications of Particle Swarm Optimization, Feb 2011.
13. Jianchao Fan,, Jun Wang, and Min Han, Cooperative Coevolution for Large-scale Optimization Based on Kernel Fuzzy Clustering and Variable Trust Region Methods,‖. IEEE Transactions on TFS., to be published.
14. Chiş, M., A new evolutionary hierarchical clustering technique, Babeş-BolyaiUniversity Research Seminars, Seminar on Computer Science, 2000,13-20.
15. Sewisy, Adel A., et al. “Fast efficient clustering algorithm for balanced data.” *Available at SSRN 2545138* (2014).
16. Lien, Li-Chuan, and Min-Yuan Cheng. “A hybrid swarm intelligence based particle-bee algorithm for construction site layout optimization.” *Expert Systems with Applications* 39.10 (2012): 9642-9650.
17. Pham D.T., Koc E., Ghanbarzadeh A., Otri S., Rahim S. and Zaidi M., “The bees algorithm - a novel tool for complex optimization problems”, In Proceedings of the Second International Virtual Conference on Intelligent Production Machines and Systems, pp.454-461, 2006.
18. ZeinEldin, Ramadan A. “A Hybrid SS-SA Approach for Solving Multi-Objective Optimization Problems.” *European Journal of Scientific Research* 121.3 (2014): 310-320.
19. Sarle, W. S., (1983) “Cubic Clustering Criterion,” SAS Technical Report A-108, Cary, NC. SAS Institute Inc.
20. Fujikawa, Y. and Ho, T. (2002). Cluster-based algorithms for dealing with missing values.In Cheng, M.-S., Yu, P. S., and Liu, B., editors, Advances in Knowledge Discovery and Data Mining, Proceedings of the 6th Pacific-Asia Conference, PAKDD 2002, Taipei,Taiwan, volume 2336 of Lecture Notes in Computer Science, pages 549–554. New York:Springer.
21. Y. Sun, Q. Zhu, Z. Chen. An iterative initial-points refinement algorithm for categorical data clustering. Pattern Recognition Letters, 2002, 23(7): 875-884
22. J. Zhang, and I. Mani, “kNN approach to unbalanced data distributions: A case study involving information extraction,” ICML’2003 workshop on learning from imbalanced datasets, 2003.
23. S.-J. Yen, and Y.-S. Lee, “Cluster-based under-sampling approaches for imbalanced data distributions,” Expert Systems with Applications, vol. 36, pp. 5718–5727, 2009.
24. Kaveh A., Laknejadi K., 2011. A novel hybrid charge system search and particle swarm optimization method for multi-objective optimization, Expert Systems with Applications, 38, pp. 15475–15488.
25. Marti, R., Laguna, M., Glover, F.: Principles of scatter search. Eur. J. Oper. Res. 169(2), 359–372(2006)
26. Yang, X.S. (2009) ‘Firefly algorithm for multimodal optimization’, SAGA 2009, LNCS 5792, pp.169–178.