

HIGH DIMENSIONAL DATA CLUSTERING THROUGH EFFICIENT EVOLUTIONARY ALGORITHM

Prof. S. Saravanan
Assistant Professor,
Master of Computer Applications,
Adhiyamaan College of Engineering,
Hosur, Tamilnadu, India

Dr. G.M. Kadhar Nawaz
Director,
Master of Computer Applications,
Sona College of Technology,
Salem, Tamilnadu, India

Abstract— Dimensionality reduction is essential in multidimensional data mining since the dimensionality of real time data could easily extend to higher dimensions. Most recent efforts on dimensionality reduction, however, are not adequate for multidimensional data due to lack of scalability. In this paper, we use the evolutionary algorithm for the dimension reduction process. Initially, our proposed evolutionary algorithm gets the values of a number of clusters and the total number of dimensions of the database. For each number of dimensions, we generate a solution. The next step of our proposed algorithm is selection of centroid for each cluster from the dataset as per the dimension of each solution. The next step is to evaluate each solution through our proposed fitness function and store the values of best one subsequently the solutions are updated through the proposed attribute evaluation process. As a consequence of the proposed evolutionary algorithm selects the best set of dimension with the best set of centroids which will apply for the K-Means clustering algorithm to create the clusters. Finally, our proposed hybrid optimization algorithm is implemented using JAVA and we compare our proposed algorithm with the standard dimension reduction algorithm in terms of clustering accuracy.

Keywords— Data Mining, Clustering, K-Means Clustering Algorithm, Evolutionary Algorithm, High Dimensional Data Clustering.

I. INTRODUCTION

It is cheering that Data mining has, of late, emerged as a promising solution for uncovering knowledge hidden in databases. It is officially described as “the non-trivial extraction of implicit, previously unknown and potentially useful information from data in databases” [6], [7] Moreover, it is extensively employed utilized for multiple needs both in the private and public sectors. Accurate employment of data mining embrace various factors such as market segmentation, fraud recognition, direct marketing, interactive marketing, market basket analysis, trend analysis and more [8]. Two common classes of data mining methods are such as descriptive and predictive. The underlying objective of descriptive data mining is to ascertain patterns, like product configurations generated in group customization applications

[9]. Certain instances of the descriptive data mining tasks are Clustering, Association rule mining and sequential pattern mining [10]. The predictive data mining is intended to create structures to determine or forecast a result such as a stock level [9]. Predictive data mining predominantly includes functions such as classification, Regression and Deviation Detection.

Clustering is the unconfirmed categorization of patterns into groups. Clustering is the function of clustering a set of objects into various subsets in such a way that objects belonging to the identical cluster are amazingly analogous to one other. Clustering is a critical function for compacting and abbreviating data with the intention of yielding a summary of the amassed data. Multidimensional data clustering is one of the widely debated areas in the arena of data mining. The wide-ranging automation and reasonable storage facilities have paved the way for mammoth quantity of data to be reachable in databases of various enterprises [5]. The crucial objective of the gigantic data compilation is the utilization of the related data to achieve aggressive recompenses, by locating formerly unknown patterns in data that can guide the course of decision making [11]. Of late, ABC algorithm [19] is found to be extensively engaged for data clustering. Afterwards, a universal stochastic clustering method [21] is brought to spotlight as a generalization of nature-inspired ant-based clustering techniques. It commences with a fundamental solution and goes on to carry out stochastic investigation to gradually fine-tune the solution until the basic clusters come to the surface, thereby leading to automatic cluster detection in datasets.

The vital issue in clustering the multidimensional arises in view of the superior dimensions of the data. The clustering procedure is thrown open to vital challenges while clustering superior dimensional data. Dimensionality reduction is one of the important keys in extracting high dimensional data. With the swift accretion of high-dimensional data like digital images, financial time series and gene expression microarrays, dimensionality reduction has emerged as a vital device for

many data extraction functions. Depending on the availability or otherwise of supervised data, modern dimensionality reduction techniques can be broadly classified into two categories such as supervised and unsupervised ones. Fisher Linear Discriminant (FLD) [12] is a fantastic pattern of supervised dimensionality reduction technique, which are competent to mine the optimal discriminant vectors during the accessibility of class labels, whereas Principal Component Analysis (PCA) [13] is a shining model of unsupervised dimensionality reduction techniques, which function by means of attempting to protect the global covariance structure of data during the period when class labels are not accessible. Of late, there is a plethora of investigations carried out in accordance with the dimensionality reduction with a view scale up the clustering of multidimensional data.

Many modern investigations have endeavored to utilize pair wise restraints or other preceding data in dimensionality reduction. Bar-Hillel et al. [14] have brightly brought to limelight the constrained FLD (cFLD) for dimensionality reduction from uniformity constraints, as an interim-measure for Relevant Component Analysis (RCA). Nevertheless, cFLD is capable of tackling the must link restraints only. Moreover, just as in FLD, cFLD faces the remarkable dilemma when restriction are narrow. Tang and Zhong [15] have efficiently employed pair wise constraints to steer dimensionality reduction, which has the prowess to utilize both must-link constraints and cannot-link constraints though it ignores the efficacy of plentiful unlabelled data. Yang et al. [16] have effectively exploited preceding data in the form of on-manifold coordinates of definite data samples for dimensionality reduction. It is clear that attaining the pair-wise constraints is generally very simple rather than getting the on-manifold coordinates of data models. Of late, [1] have deeply debated a dimensionality reduction device by means of optimization algorithms, and having got stimulated from the investigation, we have decided to launch an innovative technique for dimensionality reduction based optimization algorithm.

In this paper, we presented the proposed evolutionary algorithm for high dimensional clustering algorithm. Initially, we removed the extrinsic attributes from the original database based on our proposed fitness function. Once the extrinsic attributes are removed the original size of the dimension get reduced since the other attributes are considered as essential attributes for clustering process. From the selected set of attributes, our proposed algorithm selects the proper set of dimension with centroids to achieve the better result of the clustering algorithm. Once the process of our proposed evolutionary algorithm get completed the selected set of dimension with centroids are given to clustering algorithm to make the efficient clustering algorithm.

The rest of the paper is organized as follows: a brief review of some of the literature works in the high dimensional clustering and evolutionary algorithms in Section II and the contribution of the paper is presented in the section III. The proposed evolutionary algorithm for high dimension clustering algorithm is given in Section IV. The experimental results and performance analysis discussion are provided in Section V. Finally, the conclusions are summed up in Section VI.

II. LITERATURE REVIEW

Here, several instances of modern investigations are recounted with a view to delve deep in to the dimensionality reduction and other similar techniques related to the tackling of the dimensionality reduction dilemmas in clustering multidimensional data.

Serkan Kiranyaz et al. [1] have significantly signaled two innovative methods techniques, which fruitfully tackle various critical issues in the domain of particle swarm optimization (PSO) and hold out hopes of an incredible advance over intricate multimodal optimization menaces at superlative dimensions. The most prominent among them, known widely as the multidimensional (MD) PSO, reinvents the indigenous configuration of swarm particles so as to enable inter-dimensional passes with a dedicated dimensional PSO process.

Consequently, in the case of an MD search space, where the optimum dimension is anonymous, swarm particles is empowered to hunt for both positional and dimensional optima. This ultimately does away with the need to set a permanent dimension a priori, which is a general snag normally haunting the family of swarm optimizers. However, MD PSO is still vulnerable to untimely convergences on account of the dearth of divergence. Though there is a feast of PSO substitutes available in the literature, it is unfortunate that no one has come out successful in ensuring a vigorous solution, especially in respect of multimodal intricate issues at superior dimensions. With an eye on finding solution to the related menace, they have valiantly launched the fractional global best formation (FGBF) method, which fundamentally gathers the finest dimensional components and partially generates an artificial global best (aGB) particle that is endowed with the capacity to be a superior “guide” rather than the PSO’s native gbest particle. An wide set of tests illustrated that in both application domains, MD PSO with FGBF demonstrates an imposing speed increase and converges to the global optima at the accurate dimension in spite of the investigation space dimension, swarm size, and the complexity of the issue.

X.-S. Yang and S. Deb [2] have deftly decided to design an innovative meta-heuristic algorithm, termed as Cuckoo Search

(CS), for solving optimization problems. It is generally dependent on the obligate brood parasitic nature of certain cuckoo species in tandem with the innovative Levy flight behavior of certain birds and fruit flies. They have come out with flying colors in authenticating the new-fangled algorithm against investigation tasks and set out to contrast its efficiency with those of genetic algorithms and particle swarm optimization. In the long run, they have conducted active debates on the impact of the outcome and made reasoned recommendations for upcoming investigations.

Dimensionality reduction is one of the most important keys in mining high dimensional data. Daoqiang Zhang et al. [3] have dexterously investigated semi-supervised dimensionality reduction. In the background, in spite of numerous unlabelled models, domain knowledge in the shape of pair-wise constraints is accessible, which categorizes a couple of models into two distinct groups such as the same class (must-link constraints) or different classes (cannot-link constraints). They intelligently invented the SDDR algorithm, which is competent to maintain the intrinsic structure of the unlabelled records in addition to the must-link and cannot-link constraints explained on the labeled models in the anticipated low-dimensional space. The novel SDDR algorithm is highly proficient and possesses a closed shape solution. Test on a wide scale of data sets uphold the supremacy in respect of SDDR over several traditional dimensionality reduction techniques.

Min Soo Kim et al. [4] have majestically signaled an innovative method of an easy but efficient dimensionality reduction, called horizontal (dimensionality) reduction, for mega-size databases. Horizontal reduction transforms every text document in to a number of bitmap vectors and furnishes tight lower bounds of inter-document distances by means of those bitmap vectors. Bitmap representation is incredibly easy and awfully fast, and its instance-based nature adapts it for huge and dynamic document databases. By effectively employing the innovative horizontal reduction, we design an effective k-nearest neighbor (k-NN) investigation algorithm for text mining like classification and clustering, and its accuracy is formally established. The innovative algorithm reduces I/O and CPU overheads concurrently as horizontal reduction (1) decreases the number of accesses to documents considerably by using the bitmap-based lower bounds in sieving divergent documents well in advance, and as a result, (2) shrinks the number of CPU-intensive calculations for arriving at a real distance between high-dimensional document vectors.

Wide-ranging test outcomes underscore the fact that horizontal reduction boosts the efficiency in execution of the reduction (pre-processing) procedure by one to two orders of magnitude vis-à-vis the modern peer methods, and our innovative k-NN

search technique incredibly outclasses the classic methods by one to three orders of magnitude.

The Particle Swarm Optimization (PSO) clustering algorithm is endowed with the skills of producing further solid clustering outcomes in relation to the traditional K-means clustering algorithm. Nevertheless, at the time of clustering lofty dimensional datasets, the PSO clustering algorithm is observed to be abysmally slow in view of the fact its calculation expenses spike up tremendously with the extent of the dataset dimension. Dimensionality reduction techniques present effective solutions which enhance the evaluation interval considerably, and usher in logically correct clustering outcomes in lofty dimensional data evaluation. Xiaohui Cui et al. [17] have excellently coined an innovative investigation which blends various dimensionality reduction methods with the PSO clustering algorithm so as to shrink the intricacy of lofty dimensional datasets and thereby perk up the PSO clustering procedure. It is heartening that the innovative technique could register significant improvements in overall process period. In addition, the clustering accuracy of the dimensionality reduction PSO clustering algorithm is analogous to a different technique which exploits the whole dimension space.

Guanghui Yan [18] have got name and fame for their zooming enthusiasm and enhanced focus on the interaction between dimensionality reduction and cluster advancement in the inconstant high dimensional stream data. Based on this, they brought to spotlight the adaptive cluster evolution tracking algorithm which included the on-line fractal dimensionality reduction technique. Test outcomes emerging out of numerous authentic and artificial data sets underscore the supreme efficacy and competence of the suggested technique. In fact, discovering and tracing of cluster advancement has forever been vital to the stream data mining. In the backdrop of lofty dimensional stream data settings, it normally becomes highly complicated under the interface between dimensionality reduction and cluster evolution condition. In the foregone days, greater thrust has been placed on cluster evolution which has resulted in considerable reduction in the dimensionality space. Nevertheless, dimensionality reduction prior to the cluster evolution option is incompetent to deal with the sudden alterations which are extensively prevalent in stream data. Thus crops up the necessity for the dimensionality reduction in the course of the procedure of the cluster advancement, which is invariably well-acknowledged.

Dervis Karaboga and Celal Ozturk [19] have deftly conceived an ABC algorithm which is employed for data clustering on benchmark problems and the efficiency in accomplishment of ABC algorithm is assessed and contrasted with Particle Swarm

Optimization (PSO) algorithm and nine additional categorization methods from the literature. As much as thirteen characteristic test data sets from the UCI Machine Learning Repository are engaged to reveal the upshots of the techniques. The remarkable replication outcomes signal the fact that ABC algorithm is competent enough to be proficiently employed for multivariate data clustering. They selected the ABC algorithm, which is one of the latest hi-tech versions of the well-known optimization algorithms, mainly on account of the fact it replicates the clever foraging character of a honey bee swarm. Clustering analysis, employed in numerous domains and applications, is a very significant device and a descriptive task seeking to recognize uniform sets of objects in accordance with the values of their traits.

Bahriye Akay et.al [21] have brilliantly brought forward an innovative and improved edition of the Artificial Bee Colony algorithm, with the prime objective of effectively tackling real-parameter optimization issues. Swee Chuan Tan et.al [22] has systematically propounded a universal stochastic clustering technique which is a generalization of nature-inspired ant-based clustering method. At the outset, it yields a fundamental solution to the problem and thereafter proceeds to conduct stochastic investigation to incrementally enhance the solution till the essential clusters emerge, leading to automatic cluster detection in datasets. The significance of this technique vis-à-vis several modern techniques is that it no longer entails the users to input the number of clusters and it makes no overt postulations about the basic allocation of a dataset. Their test outcomes have upheld the supremacy in efficiency and accomplishment of the innovative technique when compared to the modern techniques in respect of clustering accuracy and effectiveness in the case of a major chunk of datasets employed in this investigation. Their theoretical analysis unequivocally demonstrated that the innovative technique exhibited linear time and space intricacies, and the pragmatic evaluation has revealed that it has the prowess to exactly and efficiently discover clusters in mega databases where the modern techniques find a waterloo and fail to deliver the goods.

With the intention of solving the constraints such as the restricted application and the inherent defects of FCM (Fuzzy C-Means) clustering algorithm, Fuhua Yu [20], has fantastically formulated an innovative and upgraded mechanical FCM clustering algorithm. At the outset, the fuzzy equivalent matrix is attained by the fuzzier by means of the standard uniform data sets. Subsequently, the objective function of the upgraded automatic FCM clustering algorithm is maximized by the amendment of membership function together with the distance measuring function. The Lagrange multiplier optimization algorithm is intended with a view to

update iteration of membership degree and clustering center. In the long run, the mechanical clustering is achieved by means of the degree of cohesion and separation. The traffic flow data of an extra-long highway tunnel in Shaanxi is considered as a real life instance for the application of the improved automatic FCM clustering algorithm. The clustering outcomes unfalteringly illustrate that fact that the validity of clustering is incredibly increased by means of the application of enhanced mechanical FCM algorithm.

III. CONTRIBUTION OF THE PAPER

In this paper, our principle objective is to achieve the best accuracy while handling the high dimensional data. In order to accomplish that goal, we provide the following contributions in this paper.

- Removing of extrinsic attributes from the high dimensional dataset thorough attribute fitness function
- Observing the appropriate set of attributes for obtaining the better clustering accuracy
- Identify the proper set of set of centroids for the selected number of attributes
- Designing the perfect fitness function to evaluate the centroids with the selected vectors

IV. PROPOSED EVOLUTIONARY ALGORITHM FOR HIGH DIMENSIONAL DATA CLUSTERING

The primary issue in the clustering of high dimensional dataset $DB = \{A_1, A_2, \dots, A_N\}$ is obtaining the better cluster accuracy. The reason behind this is contemplating of each and every vectors of the tuples for the clustering process. In order to solve this problem, in this paper, most of us decide the optimal number of attributes from the high dimensional dataset to achieving the better clustering accuracy. With the intention of selecting the essential attributes, first we evaluate the each attributes, and then we eliminate the extrinsic attributes from the original database. Once we got rid of the extrinsic set of attributes the next concern is to select the proper set of attributes and its centroid from the selected set of attributes to achieve the best accuracy of the cluster. The determination of optimal number of vectors is achieved by our proposed evolutionary algorithm also which helps to identify the exact centroids for the selected attributes. Once we establish the optimal number of attributes and its corresponding centroids by our proposed evolutionary algorithm subsequently we have given the resultant knowledge obtain from the proposed evolutionary algorithm to the well-known K-Means clustering algorithm which generates the clusters with better accuracy.

Fig. 1 represents the architecture of the proposed high dimensional clustering algorithm. Initially, our proposed algorithm selects the essential attributes from the high dimensional dataset through calculation of attribute fitness. Once the essential attributes are selected from the original database, the next step is to generate the possible solution for the proposed evolutionary algorithm. The solutions from the solution matrix are evaluated based on the proposed fitness function. From which our algorithm stores the best solution which has minimum fitness values. Then the solutions are updated randomly with the selected attributes since the searching space is defined within the selected essential attributes. This process is repeated up to N number of iterations. Ultimately the best set of attributes and its centroids are obtained as a result from our proposed evolutionary algorithm which is given to the K-Means clustering algorithm for establishing the effective partitions.

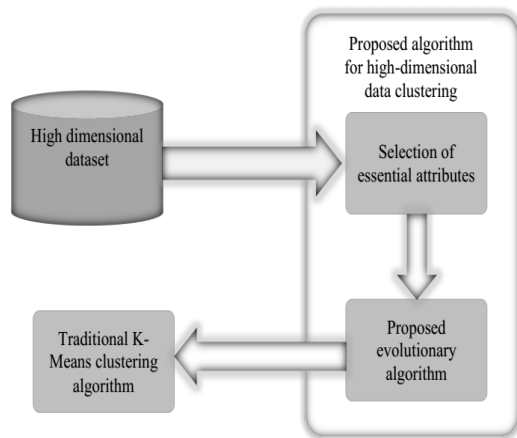


Fig 1: Represents the Architecture of the Proposed High Dimensional Data Clustering

4.1 Selection Of Essential Attributes From The Database

The process of selecting the essential attributes mainly helps to remove the extrinsic attributes of the original database, also which helps our proposed evolutionary algorithm to identify the best solution by reducing the search space of the evolutionary algorithm. Here, we evaluation process of each attribute is given below.

$$M(a_j) = \frac{1}{N} (\sum_{i=1}^N d_{p_i}) \quad (1)$$

$$V(d_{p_i}) = M(a_j) - (d_{p_i}) \quad (2)$$

$$f(a_j) = \frac{V(d_{p_i})}{N} \quad (3)$$

The above mentioned equation1, used to determine the mean value of each attribute and the equation 2 is used for calculating the variance value of each data point in an attribute. Finally, the equation (3), represents the fitness value of the

attribute. With the intention of selecting the best set of attributes, in this paper, we establish the user defined threshold for the fitness value of the attributes. If the fitness value of the attribute is greater than the user defined threshold value, then those attributes are considered as essential attributes for that database.

4.2 The Proposed Evolutionary Algorithm

Our proposed evolutionary algorithm employed for selecting the best set of attributes with the centroids. We designed the effective fitness function in order to evaluate the solutions of the proposed algorithm. The process of our proposed evolutionary algorithm is applied under the following.

- Generation of solution matrix for each number of dimension
- Evaluation of solutions through fitness function
- Selection of best solutions
- Updation of solution matrix

4.3 Generation of Solution Table for Each Number of Dimension

The high dimensional database consists set of attributes $\{A_i\}$ where $(1 \leq i \leq M)$. The principle issue in the clustering of high dimensional dataset is acquiring the better cluster accuracy. The explanation for this really is taking in top consideration of every vector of the tuples for the clustering process. With the aid of selecting the best set of attributes, in this paper we select and evaluate the different set of attributes. The selection of the best set of attributes is performed by the proposed evolutionary algorithm. By offering the different set solutions to evolutionary algorithm which will produce the best set of attributes for better cluster accuracy. The best set of attributes is selected by the proposed fitness function which affects the clustering accuracy. As a way to select the best solution, the value of fitness of the solution must be less.

In this paper, we have given the set of input solutions to the evolutionary algorithm. Each result has a different number of dimensions which helps to identify the best centroids in each solution. The selection of number of attribute has the range value; minimum to maximum will be calculated as $R(D) = (2 \text{ to } M - 1)$ where M is the maximum dimension of the dataset. For that purpose initially, we generate the solution matrix which holds the number of solutions S_1, S_2, \dots, S_R where R is the maximum number of solutions which is calculated by $R(S) = (R(D) - 1)$. The following table 1 represents the solution matrix.

Table 1 : Solution Matrix

Solutions	D ₁	D ₂	D ₃	D _m	D _{M-2}	D _M	D _M
S ₁	1	0	0	0	0	0	1
S ₂	0	1	1	0	1	0	0
S _r	1	0	0	1	0	1	1
S _{R-1}	1	1	1	1	1	0	0
S _R	1	1	1	0	1	1	1

Table 1: represents the solution matrix From the above solution matrix (Table D), where the value of 1 indicates the selected attributes for the evaluation process, and the 0 indicates neglected attributes of the solutions. Once the generation of solution matrix gets over, the following step is to evaluation process of each solution which will be done by the proposed fitness function. In order to evaluate the solution, we utilized the K-Means clustering algorithm as an initial clustering process. With the help of outcome result of the K-Means algorithm.

4.4 K-Means Clustering Algorithm

The K-Means algorithm is applied as initial clustering for each solution. For each solution, the K-Means clustering algorithm, considers the database with the selected attributes of each solution. Once the database selected, the K-Means clustering algorithm makes the K number of partitions. The procedure of the K-Means clustering algorithm is given below.

The K- Means clustering algorithm initiates its operation by obtaining the value of K from the user, then it selects the K number of centroids from the database then it estimates the distance between each centroids with other data points for making the partition. The Euclidean distance function is applied for estimating the distance value. The Euclidian function is given in the following equation (4).

$$D_{i,j} = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \dots + (x_n - y_n)^2} \quad (4)$$

The following table II represents the distance matrix, which interprets the distance value of each data point with every cluster centroid. Once the distance matrix generated, then the clusters matrix is generated based on the distance between them which is represented in the following table 3.

Table 2: Distance Matrix

Data point / cluster centroids	Centroid C ₁	Centroid C _j	Centroid C _K
D ₁	d ₁₁	d _{1j}	d _{1K}
D ₂	d ₂₁	d _{2j}	d _{2K}
D _i	d _{i1}	d _{ij}	d _{iK}
D _{N-1}	d _{N-11}	d _{N-1j}	d _{N-1K}
D _N	d _{N-11}	d _{N-1j}	d _{N-1K}

Table 3: Cluster Matrix

Data point / cluster centroids	Centroid C ₁	Centroid C _j	Centroid C _K
D ₁	0	0	1
D ₂	1	0	0
D _i	1	0	0
D _{N-1}	0	1	0
D _N	0	0	1

Table III represents the cluster matrix. The K-Means algorithm evaluates the distance value of a single data point with regard to each centroid. The data point belongs to the centroid which has minimum of distance value among them. From the above mentioned table 3, the value 1 indicates the minimum value distance of each data point.

4.5 Updation of Centroids

Each centroid has the number of data points, which is employed for the centroid updation process. The updation process of the centroid is achieved by updating the each attribute of the centroids. The following equation (5) is used for updating the attribute of centroids.

$$C_j(a_i) = \frac{1}{n} \sum_{i=1}^n (a_i^j) \quad (5)$$

From the above equation 1, where C_j(a_i) represents the updation of attribute from the centroid a_i from the centroid C_j. The value of ‘n’ represents the number of data point present in the cluster C_j.

Once the centroids of the solutions get updated, the distance calculation process gets repeated until the centroids of the attribute become same in consecutive iterations. Finally, the each cluster has the same data point.

4.6 Calculation of fitness function of solutions

We determine the proposed fitness function to evaluate the solution. We analyzed the each solution with its resultant clustered data. The estimation of the fitness function is exemplified in the following equation (6). The upcoming table 4 represents the evaluation matrix (sample outcome) of the clustering algorithm.

Table 4: Evaluation Matrix

Data point / cluster centroids	Centroid C ₁	Centroid C _j	Centroid C _K
D ₁	0	0	d _{1K}
D ₂	d ₂₁	0	0
D _i	d _{i1}	0	0
D _{N-1}	0	d _{N-1j}	0
D _N	0	0	d _{N-1K}
Avg	Avg (d ₁)	Avg (d _j)	Avg (d _K)
Max	Max (d ₁)	Max (d _j)	Max (d _K)

$$F(S_r) = \left(\frac{\sum_{i,j=1}^{i=N_r, j=N} \binom{d_{ij}}{N_j}}{\sum_{i,j=1}^{i=N_r, j=N} \max(d_{ij})} \right) \times \left(\frac{\sum C_k}{K} \right) \quad (6)$$

From the above equation (6), where 'i' indicates the data point id and the 'j' indicates centroid id and N_j indicates the total number of data points of cluster based on the centroid j. $\max(d_{ij})$ represents the maximum value of the distance of cluster based on the centroid j. The symbol C_k indicates the distance between the centroids and K represents the total number of clusters.

4.7 Selection of Best Solution

Our proposed fitness function is best when it became less, since we store the best solutions in each iteration. After finishing the final number of iterations we select the best of the best set of solution from the history of the best solution as we laid earlier.

4.8 Updation of Solutions

In the initial process of the proposed algorithm, we removed the extrinsic attributes from the original database since all the others are considered as important attributes. Since we can choose any of the attributes randomly for updating the existing solution, the removal of extrinsic attribute process, helps the evolutionary algorithm in terms of reducing the search space. Since our proposed algorithm selects the best set of attributes in order make the cluster efficiently.

V. EXPERIMENTAL RESULTS AND PERFORMANCE ANALYSIS DISCUSSION

The experimental results of the proposed approach for mining relational patterns are presented in this section. The proposed approach has been implemented using JAVA and the performance of the proposed system is analyzed using the accuracy and running time. With the help of referencing [23] we calculate the accuracy of the clustering algorithm for each dataset in the benchmark datasets [24]. The running time is evaluated by how much time (seconds) the proposed system takes for clustering the each dataset in the benchmark.

5.1 Experimental Environment and Dataset

We have practiced the real dataset for mining relational patterns and experimentation is performed on a 3.0 GHz dual core PC machine with 2 GB main memory. We have utilized the real dataset called time series benchmarks which having 9 time series datasets [24] has been compiled to access the time series clustering and clustering algorithms and the information about the dataset is given in the following table 1 which consists of a number of classes of each time series dataset and the size of each dataset including training and testing and length of the dataset.

Table 5. Dataset Details

Dataset	Number of class	Size of dataset	Length
CBF	3	1200	128
ECG 200	2	200	96
Face Four	4	112	350
Gun-Point	2	200	150
Lighting2	2	121	637
Lighting7	7	143	319
OSU Leaf	6	442	427
Trace	4	200	275
Yoga	2	3300	426

In the previous paper [25] high dimensional clustering algorithm is compared with many existing algorithms in order to prove the efficiency of the algorithm. Also the result of the paper [25], represents their algorithm is performed well when compared with the existing techniques in terms of clustering accuracy. Since, in this paper, we compare our proposed high dimensional clustering algorithm with the existing clustering algorithm in terms of accuracy.

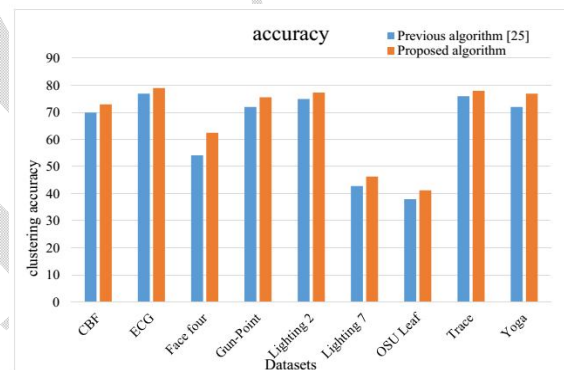


Fig 2 : Evaluation of Accuracy in Different Datasets

By analyzing the above Fig. 2, in every datasets, our proposed algorithm performed well than the algorithm [25] in terms of clustering accuracy. The main reason behind this is we designed effective fitness function for evaluating the solutions from the evolutionary algorithm also we minimize the searching space of the proposed algorithm by removing the extrinsic attributes of the database as a result the dimension get reduced automatically. Moreover, our proposed algorithm selects the best set of attributes from the selected essential attributes. Our proposed algorithm performed 15.39% than the algorithm [25] in face four dataset this is the maximum performance of our proposed algorithm and it achieved minimum performance 2.6% than the algorithm [25] in trace dataset.

VI. CONCLUSION

We proposed an efficient algorithm for high dimensional data clustering. Initially, our proposed algorithm selected the

essential attributes from the high dimensional dataset through calculation of attribute fitness. Once the essential attributes were chosen from the original database, afterwards it generated the possible number of solutions for the proposed evolutionary algorithm. The results from the solution matrix are evaluated based on the proposed fitness function. From which our algorithm retains the best solution as which has minimum fitness values. Then the solutions are updated randomly with the selected attributes since the searching space is limited within the selected essential attributes. This process is repeated up to N number of iterations. Ultimately the best set of attributes and its centroids are obtained as a consequence from our proposed evolutionary algorithm which is given to the K-Means clustering algorithm for establishing the effective partitions. Finally, our proposed algorithm was implemented using JAVA and we compared our proposed algorithm with the previous high dimensional data clustering algorithm [24] in terms of clustering accuracy and we proved our proposed clustering algorithm was performed well in terms of clustering accuracy.

References

- [1] Serkan Kiranyaz, Turker Ince, Alper Yildirim, and Moncef Gabbouj, "Fractional Particle Swarm Optimization in Multidimensional Search Space", IEEE Transactions On Systems, Man, And Cybernetics, VOL. 40, NO. 2, pp: 298-319, 2010.
- [2] X.-S. Yang, S. Deb, "Cuckoo search via Levy flights", in: Proceedings Of World Congress on Nature & Biologically Inspired Computing (NaBIC 2009), December 2009, India. IEEE Publications, USA, pp. 210-214 (2009).
- [3] Daoqiang Zhang, Zhi-Hua Zhou, Songcan Chen, "Semi-Supervised Dimensionality Reduction", In Proceedings of the 7th SIAM International Conference on Data Mining, pp: 11–393, 2007
- [4] Min Soo Kim , Kyu-Young Whang and Yang-Sae Moon, "Horizontal Reduction: Instance-Level Dimensionality Reduction for Similarity Search in Large Document Databases", IEEE International Conference on Data Engineering, pp: 1061 – 1072, 2012.
- [5] Vaibhav Kant Singh, Vijay Shah, Yogendra Kumar Jain, Anupam Shukla, A.S. Thoke, Vinay Kumar Singh, Chhaya Dule and Vivek Parganiha, "Proposing an Efficient Method for Frequent Pattern Mining", World Academy of Science, Engineering and Technology, vol.61, pp. 384-390, 2008.
- [6] Osmar R. Z., "Introduction to Data Mining", In: Principles of Knowledge Discovery in Databases. CMPUT690, University of Alberta, Canada, 1999.
- [7] Kantardzic, Mehmed. "Data Mining: Concepts, Models, Methods, and Algorithms", John Wiley and Sons, 2003.
- [8] E. Wainright Martin, Carol V. Brown, Daniel W. DeHayes, Jeffrey A. Hoffer and William C. Perkins, "Managing information technology", Pearson Prentice-Hall 2005.
- [9] Andrew Kusiak and Matthew Smith, "Data mining in design of products and production systems", in proceedings of Annual Reviews in control, vol. 31, no. 1, pp. 147- 156, 2007.
- [10] Mahesh Motwani, J.L. Rana and R.C Jain, "Use of Domain Knowledge for Fast Mining of Association Rules", in Proceedings of the International Multi-Conference of Engineers and Computer Scientists, 2009.
- [11] Satheesh, A. Mishra, D.K. and Patel, R. "Classification Rule Mining for Object Oriented Databases: A Brief Review", in proceedings of the First International Conference on Computational Intelligence, Communication Systems and Networks, pp. 259 - 263, Indore, July 2009.
- [12] R. A. Fisher, The use of multiple measurements in taxonomic problems, Annals of Eugenics, 7 (1936), pp. 179–188.
- [13] I. Joliffe, Principal Component Analysis, Springer, New York, NY, 1986.
- [14] A. Bar-Hillel, T. Hertz, N. Shental, D. Weinshall, Learning a mahalanobis metric from equivalence constraints, Journal of Machine Learning Research, 6 (2005), pp. 937–965.
- [15] W. Tang and S. Zhong, Pairwise constraints-guided dimensionality reduction, in SDM'06 Workshop on Feature Selection for Data Mining, Bethesda, MD, 2006
- [16] X. Yang, H. Fu, H. Zha, J. L. Barlow, Semisupervised nonlinear dimensionality reduction, in ICML'06, Pittsburgh, PA, 2006, pp. 1065–1072.
- [17] Xiaohui Cui , Beaver, J.M. ; St. Charles, J. ; Potok, T.E. , "Dimensionality reduction particle swarm algorithm for high dimensional clustering ", Swarm Intelligence Symposium, pp:1-6, 2008.
- [18] Guanghui Yan, "Integrating fractal dimensionality reduction with cluster evolution tracking", 2011 Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), vol. 3, pp. 1668 - 1672, 2011.
- [19] DervisKaraboga, CelalOzturk, "A novel clustering approach: Artificial Bee Colony (ABC) algorithm", Applied soft computing, vol. 11, pp: 652-657, 2011
- [20] Fuhua Yu, "An Improved Automatic FCM Clustering Algorithm", 2010 2nd International Workshop on Database Technology and Applications (DBTA), pp. 1 - 4 , 2010.
- [21] BahriyeAkay ,DervisKaraboga, "A modified ABC algorithm for real parameter optimization", information sciences, vol. 192,June ,2012.
- [22] SweeChuanTan,Kai Ming Ting,ShyhweiTeng, "A general Stochastic clustering method for automatic clustering Discovery", pattern recognition, vol.44,October, 2011.
- [23] Yun Yang and Ke Chen, "Temporal Data Clustering via Weighted Clustering Ensemble with Different Representations", IEEE Transactions on Knowledge and Data Engineering, vol. 23, no. 2, pp. 307-320, 2011.
- [24] E. Keogh, Temporal Data Mining Benchmarks, http://www.cs.ucr.edu/~eamonn/time_series_data, 2010.
- [25] S.Saravanan and G.M.Kadhar Nawaz, "Ensemble-Based Time Series Data Clustering for High Dimensional Data", International Journal of Innovative Computing, Information and Control, vol.10, no. 4, pp.1457-1470, 2014.