# HYBRID OPTIMIZATION BASED CLUSTERING APPROACH FOR EFFECTIVE PREDICTING AND RECOMMENDATION OF MOVIES

P JAYAPRIYA[1]
V DURGA DEVI[2]
Dr R PRIYA[3]
1. Assistant Professor,
Department of Computer Science,
Ethiraj College for Women,
Chennai.
2. Research Scholar , Department of Computer Applications, VISTAS,Pallavaram,.
3. Professor, Department of Computer Applications, VISTAS,Pallavaram,

## ABSTRACT

Fast and high feature clustering algorithms shows an important role in data mining for active movie recommendation , summarize, and organization of data. The K-Means algorithm is the supreme universally used partitioned clustering algorithm since it can be simply implemented and is the supreme proficient in terms of the implementation time. This paper proposes two methods for efficient document clustering; these proposed methods concerning the application of computing methodology as an intellectual enhanced genetic based algorithm.

***Keywords:*** *Cluster Centroid, Optimization, K-Means clustering, Particle Swarm Optimization, Genetic Algorithm*

## INTRODUCTION

Because of revolution in entertainment industry, source of entertainment has been increasing rapidly in today’s world. Users have to choose entertainment products from a vast amount of options which can be overwhelming for any user. As a result, recommendation systems for any product have gained popularity for every field in digital systems. As the entertainment world is booming with huge amount of data, automated recommendation systems having different approaches can be useful for recommending products to users. In this paper, we have proposed a machine learning based way to recommend movies using clustering and machine learning approaches. The objective of this paper is to separate users using clustering algorithm in order to find users with similar taste of movies. Machine learning approaches are used to guess what rating a particular user might give to a particular movie so that this information can be used to recommend movies to viewers. In this recommendation system, we used the publicly available data from MoveLens [1].

In general, a movie recommendation system compares user’s profile or usage data to some reference characteristics and combines the user’s social environment to make movie recommendations. This type of recommendations is based on user. However, this type of recommendations may not work or make inaccurate recommendation in the following situations. The user does not have strong profile setting in the system. There are many users who do not want to set their profile due to laziness or privacy concerns[3]. In this case, most recommendation systems consider the user's social connection in the system such as friends, classmates, families, and colleagues. However, the tastes of the movies may be various even among best friends. What is worse, the "friends" that the user has added in the website may not be people with the same interest. For example, in a community network or local network such as a university or college, like our experimental environment, users' social connections are built mainly because they are in the same university with the same major and similar age; however, their tastes towards movies may be totally different, which fails the fundamental bias in the recommendation system mentioned above. Because of the situation stated above, we propose the use of machine learning (ML) techniques (clustering) to analyze the movie features and system logs (user's voting logs) to make correct recommendations more adequately. In this proposed approach, we compute a distance matrix for the movie features and apply the clustering techniques to classify movies into different areas off-line. For every user logged on the system, we recommend movies from the clusters combined with the user's majority voting result in real time.In order to compare the accuracy and efficiency, we have implemented different clustering techniques as follows: DBSCAN (density-based clustering), affinity propagation, hierarchical clustering, and random clustering as a base line[4].In the MovieLens dataset, there are two files containing information about movies and users. The movie data has been split based on their genre and later outer joined with ratings of movies in order to get user preference, average rating and consumption ratio for each genre of movies in three separate approaches. This resulted in each tuple having 18 attributes in all the approaches since there are 18 genres listed in MovieLens[1] dataset. Clustering was used to separate dissimilar users and the result was compared in the three approaches to choose the best one. Principal Component Analysis [2] (PCA) was used to decrease the dimension for a better clustering result. It was used in such a way that we get don't lose any user data after clustering. Finally, the rating was included in the last column as output column which was later on used for the neural network

## 2.RELATED WORKS

H. Chen and A. Chen [5] designed a music recommendation system. Their workflow consists of analyzing the music objects, determining the representative track, extracting six features from the track. That is how music objects are grouped. In order to understand the interests of the users, their access history is analyzed. Recommendation methods are proposed mainly based on the preferred degrees of the listeners to the music groups.

Ahmed, et al. [6] demonstrated how TV series recommendation can be challenging and different than movie recommendation. In TV series recommendation, time commitment issue needs to be analyzed Other than analyzing genre, which adds some extra work on TV series recommendation and this paper showed a way to achieve that using fuzzy systems.

Park, Hong, and Cho [7] proposed a recommendation system which is personalized where users' preference is reflected by Bayesian Networks. The parameters are learned from a dataset whereas the structure of the Bayesian Network was built by an expert. The system they proposed works by collecting context information such as location, time, and weather condition. It also analyzes user request from the mobile device to infer the most favored item so that it can provide an appropriate facility by showing it in the map.

Huang and Jeng [8] worked on audio recommendation system. Their system takes user assigned rating for songs of his/her song collection and extracts the audio signature. LBG vector quantization is used by the system to rate new audio file.

S. Kalyani, et al., [9] Safety assessment is a major concern in the planning and operation of a power system studies. Conventional method of evaluating the security Played by computer simulation implies long and generates large result. Secure / insecure under given operating condition and contingency this article presents a K-means approach for ranking the states of the power system. The simulation results of the algorithm are compared with the K-means clustering without supervision proposed, which use different methods for initializing the cluster center.

R.J. Kuoet al., [10] although cluster analysis algorithms are constantly improving, most clustering algorithms has yet to define the number of clusters. Thus, this study proposes a dynamic clustering approach based on the novel particle swarm optimization (PSO) and Genetic Algorithm (GA) (DCPG) algorithm. The proposed algorithm DCPG Data can be automatically ammunition by examining the data without a number of pre-specified clusters. The results of calculation of four reference data sets indicate that the algorithm has better DCPG validity and stability of the dynamic clustering approach based on binary PSO (DCPSO) and dynamic clustering approach based on GA (DCGA) algorithms.

S. Rana et al., [11] the data grouping is the most popular data analysis method in data mining. It is the method that the parties of the data object to significant groups. It has been applied in many fields such as image processing, pattern recognition and learning of the machine where the data sets are many shapes and sizes. This article presents a new improved algorithm named Adaptive Boundary small particles Swam Optimization (BR-APSO) algorithm with the limit restriction policy.

## 3. PROPOSED METHOD

Clustering is the development of distinguishing usual federations or clusters in multidimensional statistics established on some similarity processes . Detachment dimension is normally castoff for estimating similarities in the middle of patterns. In specific the delinquent is detailed as follows: assumed N objects, assign each entity to some of K clusters and minimize the summation of squared Euclidean detachments in the middle of each entity and the midpoint of the cluster fitting to each such assigned object.Benevolences K-means clustering methodology for control structure safety grouping. Proposed algorithm syndicates particle swarm optimization by means of old-fashioned K-means algorithm.

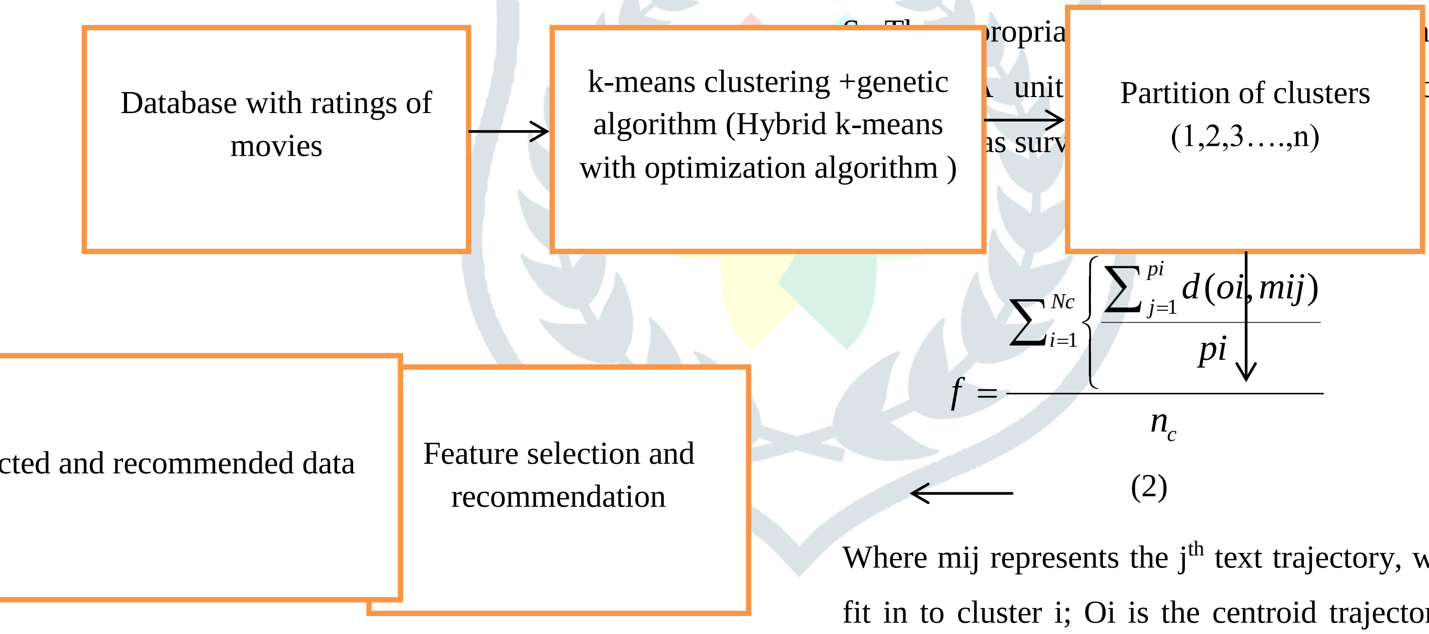
Database with ratings of movies
k-means clustering +genetic algorithm (Hybrid k-means with optimization algorithm )
Partition of clusters (1,2,3….,n)
Feature selection and recommendation
cted and recommended data

*Figure-1 Blocked diagram of proposed method*

### 3.1 Proposed K Means + Genetic Algorithm

This research paper proposes two hybrid performances for clustering typescript in database of movie ratings: Hybrid of K-Means algorithm and Genetic algorithm.

The hybrid of K-Means and GA is projected to be primed with K-Means unit and then GA is functional on the preliminary consequences produced by K-Means unit. In K-Means unit the decontrol of the cluster centroid is finished as .

$$c_j = \frac{1}{n_j} \sum_{\forall d_j \in s_j} d_j$$

(1)

Where $d_j$ represents the document trajectories that fit in to cluster $S_j$; $C_j$ standpoints for the centroid vector; $n_j$ is the sum of followers fit in to cluster propria alize unit ch is as surv

$$f = \frac{\sum_{i=1}^{Nc} \left\{ \frac{\sum_{j=1}^{pi} d(oi,mij)}{pi} \right.}{n_c}$$

(2)

Where mij represents the $j^{th}$ text trajectory, which fit in to cluster i; Oi is the centroid trajectory of the ith cluster; d(oi,mij) is the detachment the middle of document mij and the cluster centroid Oi.; pi standpoints for the total database of movie ratings, which fit in to cluster Ci; and Nc standpoints for the quantity of constellations.

The Pseudo code for K-Means + Genetic Algorithm encompasses of the subsequent steps:

**Step 1:** Select K-points as preliminary centroids

**Step 2:** Repeat

a) Practice K-clusters by transmission for each fact to its neighboring centroid.

b) Recalculate the centroid of separate cluster.

**Step3:** Up to centroid doesn't alteration

**Step 4:** programming A on primary clusters produced through K-Means

a) Modify the Constituent part (Clusters)

b) Modify Vi(t), Vmax, c1 and c2

c) Modify Inhabitants magnitude and repetitions

d) Modify clusters to input data

e) Attain the innovative point

The objective in this circumstance is to develop the extrapolation competence of the prototypical by creating it capable to envisage the TBVH of MH while it still transfers in the analysis of A. As the MH is also distant from the BBS to become a realistically correct significance of $b$, we initially necessity to bargain this distance and the angle $\beta$ in edict to compute distance $z$.

$$c^2 = d^2 + b^2 - 2db\cos\theta$$

Consequently,

$$\theta = \cos^{-1}\left(\left(c^2 - d^2 - b^2\right) / 2db\right) \tag{3}$$

Dependent on which crosswise of line AB point X lies,

$$\text{Angle } \beta = |x - \theta| \tag{4}$$

In view of triangle BYC, we consume

$$t = b\cos\beta \tag{5}$$

$$y = b\sin\beta \tag{6}$$

**Step 5:** Recapitulate Genetic

a) Discover the endearing points

b) Modernize Velocity and Locus

**Step 6:** Estimate the power of Genetic

a) Recapitulate Generation

b) Consume weak constituent part

c) Recalculate the locus

**Step 7:** Exit when the extreme number of repetitions satisfied or any other ending principles is stretched.

Consequently, in triangle BYX,

$$s = \sqrt{r^2 - b^2\sin^2\beta} \tag{7}$$

$$z = b\cos\beta + \sqrt{r^2 - b^2\sin^2\beta} \tag{8}$$

Therefore the TBVH module for this state is,

$$TBVH = \frac{b\cos\beta + \sqrt{r^2 - b^2\sin^2\beta}}{v} \tag{9}$$

This is related to the equation attained in (4).

### 3.3 Feature Selection for recommendation

Database of movie ratings essential to be signified in an appropriate method for clustering. The supreme communal illustration contains the Vector Space Model (VSM) which gives database of movie ratings as a bag-of-words and practices words as a quantity to observe similarity among database of movie ratings. In this prototypical, individual document $D_i$ is positioned as an argument in m dimensional vector space,

$$Di = (w_{i1}, w_{i2}, ......, w_{im}), i = 1, ...., n \qquad (10)$$

Where the measurement is the equivalent as the number of terms in the document group. For each module of such a direction replicates a term inside the specified document. The significance of separate module rest on the grade of connection in the middle of its related term and the corresponding document. The supreme mutual term weighting pattern to extent these interactions is the Term Frequency (tf) and tf-idf. The tf-idf is computed as lower than

$$w_{ij} = n_{ij} \times \log\left(\frac{n}{n_j}\right) \qquad (11)$$

Where:

$n_{ij}$is the tenure frequency (i.e., signifies several term $T_j$ take place in document Di),

nj represents the numeral of database of movie ratings in which tenure $T_j$ performs.

The tenure log ($n/n_j$) is the idf feature and interpretations for the universal weighting of tenure $T_j$.

## 4.EXPERIMENTAL RESULTS

In this sector, fine points of the total results of the proposed algorithm are discussed. This segment has been separated into binary subsections. Initially the functioning of the proposed structure and modification in the cluster centers is exemplified. In addition to estimate the enactment of the proposed clustering algorithm, a small number of experiments have been accompanied on two artificial engendered data set complications and additional two with standard data mining benchmark algorithms.

GA is applied initial to the data set and the acquired results are made known in Figure 3 for dual cluster origination. The acquired results as of GA are then deal with over K-Means algorithm for supplementary filtering the cluster establishment. Figure 2displays the operational of proposed Hybrid clustering algorithm. It be able to see that the cluster midpoints are promote shifted. It can similarly realize that the midpoints are moving further in the direction of the centre of cluster and mutually midpoints are moving distant from each other i.e. intensification of distance among the cluster centers.

.

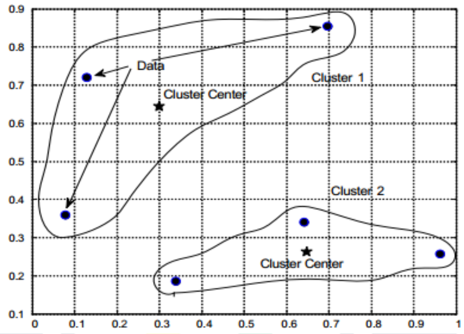

*Figure2. Cluster Centre and cluster formation by GA*

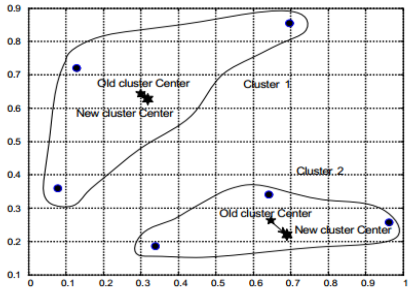

*Figure3. Cluster Centre refinement by K-Means Algorithm*

The accurateness and toughness of our proposed algorithm have been verified on three dissimilar difficulties. The classification difficulties are as follows:

**(i) problem I:** This problematic formulation is completing consistent with subsequent classification rule

**(ii) problem II:** The problem is stimulating in that only one of the inputs is certainly pertinent to the establishment of the programs

**(iii) Clustering Data Set:** This is conceivably the greatest recognized database to be establishing the pattern recognition fiction. The foremost determination of our proposed Hybrid Sequential clustering algorithm is to equate the superiority of the particular clustering, where superiority is restrained rendering to the following three standards:

The quantization error by way of distinct in equation (14)

$$Q_e = \frac{\sum_{j=1}^{c}\left[\frac{\sum d\left(\overline{X^p}, z^j\right)}{N_0}\right]}{C} \qquad (14)$$

Where $d\left(\overline{X^p}, z^j\right)$ detachment to centroid, N$_0$ is number of database of movie ratings trajectories to be clustered, c is the number of cluster to be molded.

The intra-cluster distances, i.e. the distance among database of movie ratings trajectories inside a cluster, where the objective is to minimalized the intra-cluster distances and is assumed

$$Intra = \frac{1}{n}\sum_{j=1}^{c} \| \overline{X^j} - z^j \|^2 \qquad (15)$$

The inter-cluster distances, i.e. the distance among the centroids of the clusters, where the objective is to make the most of the distance among clusters is assumed by equation (16)

$$Intra = \min(\| z^i - z^j \|^2 \qquad (16)$$

The Experimental results attained from the five clustering algorithms are summarized in below Tables 1 to 3.The eccentricities in the results attained are minimalized in the projected algorithm. All other algorithms have improved result in single or another instance but there is no consistency in the result attained. It is only the proposed K-Means Genetic Algorithm which makes superlative between them.

**Table 1.Comparison of K-Means, PSO, Hybrid, and Hybrid Sequential clustering algorithm,K-Means Genetic Algorithm with Quantization Error**

| Algorithm | Problem I | Problem II | Clustering Data Set |
|---|---|---|---|
| K-Means Algorithm | 0.984 | 0.264 | 1.139 |
| Particle Swarm Optimization Algorithm | 0.769 | 0.252 | 0.774 |
| Hybrid | 0.768 | 0.250 | 0.633 |
| Hybrid Sequential clustering algorithm | 0.772 | 0.260 | 0.982 |
| K-Means + Genetic algorithm | 0.764 | 0.250 | 0.628 |

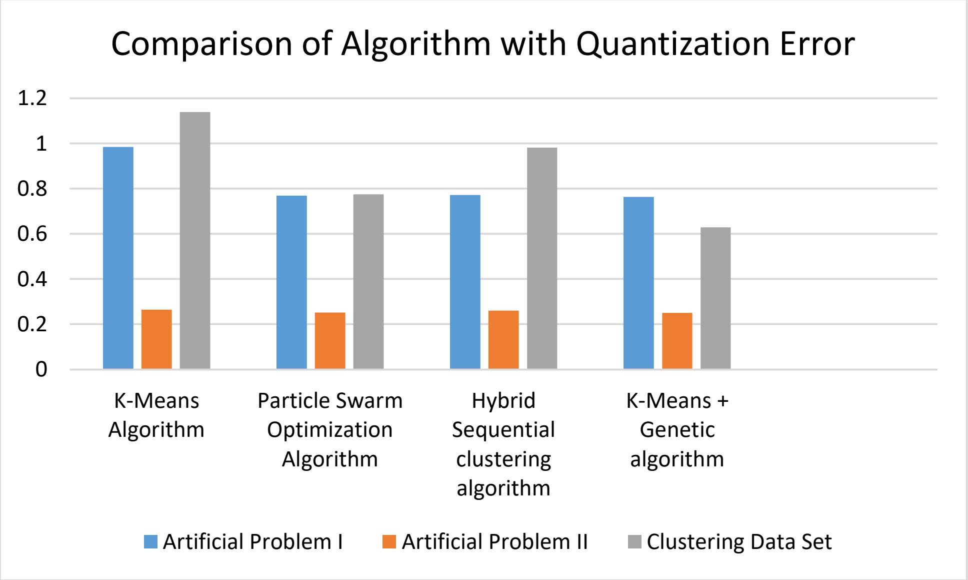

*Figure4. Comparison of Algorithm with Quantization Error*

**Table 2.Comparison of K-Means, PSO, Hybrid, and Hybrid Sequential clustering algorithm, K-Means Genetic Algorithm with Intra-cluster Distance**

| Algorithm | Artificial Problem I | Artificial Problem II | Clustering Data Set |
|---|---|---|---|
| K-Means Algorithm | 3.678 | 0.911 | 3.374 |
| Particle Swarm Optimization Algorithm | 3.826 | 0.873 | 3.489 |
| Hybrid | 3.823 | 0.869 | 3.304 |
| Hybrid Sequential clustering algorithm | 3.892 | 0.899 | 3.378 |
| K-Means + Genetic algorithm | 3.647 | 0.864 | 3.300 |

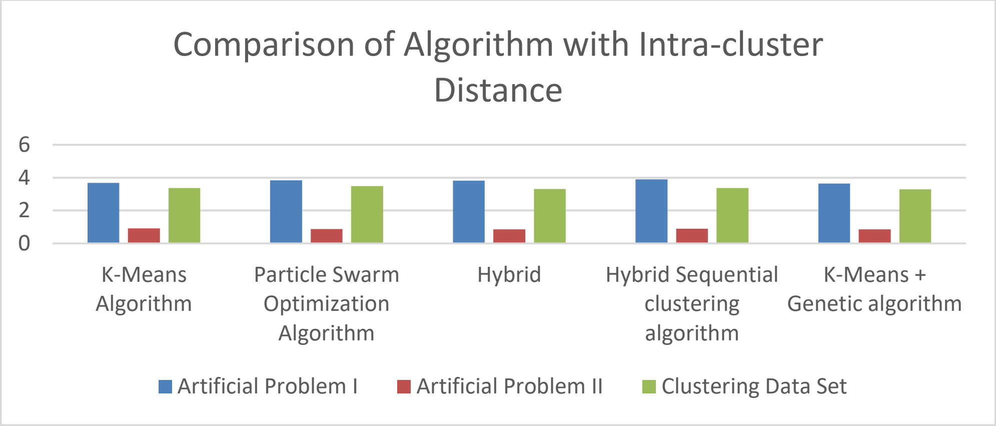

*Figure5. Comparison of Algorithm with Intra-cluster Distance*

**Table 3.Comparison of K-Means, PSO, Hybrid, and Hybrid Sequential clustering algorithm, K-Means Genetic Algorithm with Inter-cluster Distance**

| Algorithm | Artificial Problem I | Artificial Problem II | Iris Data Set |
|---|---|---|---|
| K-Means Algorithm | 1.771 | 0.796 | 0.887 |
| Particle Swarm Optimization Algorithm | 1.142 | 0.815 | 0.881 |

| | | | |
|---|---|---|---|
| Hybrid | 1.151 | 0.814 | 0.852 |
| Hybrid Sequential clustering algorithm | 1.151 | 0.815 | 0.863 |
| K-Means + Genetic algorithm | 1.779 | 0.815 | 0.894 |

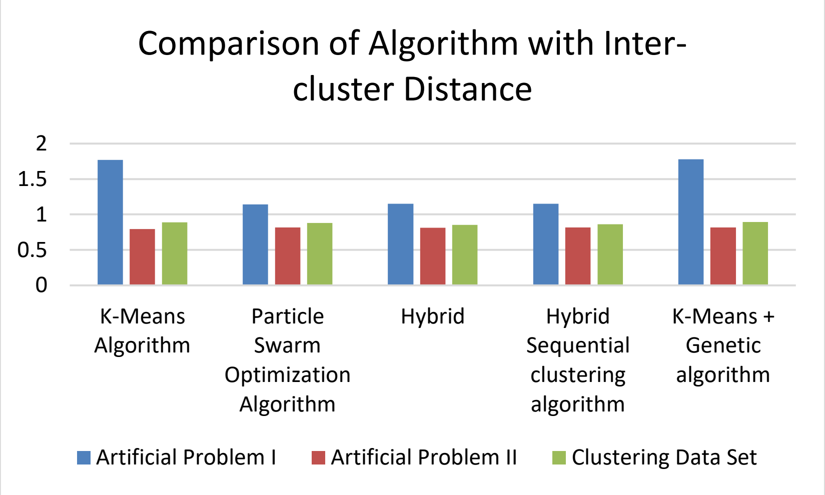

***Figure6. Comparison of Algorithm with Inter-cluster Distance***

Table 1, Table 2 and Table 3 contemporary the comparison of algorithms allowing for intra- and inter-cluster distances. These limitations are deliberated to guarantee compact clusters with slight deviance from the cluster centroids and grander parting among the dissimilar clusters. It can be realized from the consequences that K-Means Genetic algorithm effectively attain improved results than its corresponding item.

## DISCUSSION AND CONCLUSION

In this research paper, K-Means algorithm, which is an original, modest and robust optimization method, is cast off in clustering of the standard classification complications for classification purpose. Five algorithms are verified, that is to say atypical K-Means, PSO, K-Means Genetic algorithm, Hybrid method and the Hybrid Sequential clustering algorithm, wherever the swarms novelty the clusters centre and additional filtering attained over K-Means algorithm. The enactment of the K-Means algorithm is equated with Particle Swarm Optimization algorithm and additional methods which are extensively used by the investigators.

## REFERENCE

[1] “MovieLens.” GroupLens, 18 Oct. 2016, grouplens.org/datasets/movielens/.

[2] “Principal component analysis.” Wikipedia, Wikimedia Foundation, 2 Sept. 2017, en.wikipedia.org/wiki/Principal_component_analysis

[3] M. A. Ghazanfar and A. Prügel-Bennett, “An improved switching hybrid recommender system using Naive Bayes classifier and collaborative filtering,” in Proceedings of the International MultiConference of Engineers and Computer Scientists (IMECS '10), pp. 493–502, March 2010.

[4] A. Bellogín, I. Cantador, P. Castells, and A. Ortigosa, “Discovering relevant preferences in a personalised recommender system using machine learning techniques,” in Proceedings of the ECML-PKDD Workshop on Preference Learning, 2008.

[5] Chen, Hung-Chen, and Arbee L. P. Chen. “A music recommendation system based on music data grouping and user interests.” Proceedings of the tenth international conference on Information and knowledge management - CIKM01, 2001, doi:10.1145/502585.502625.

[6] Ahmed, Muyeed, et al. “TV Series Recommendation Using Fuzzy Inference System, K-Means Clustering and Adaptive Neuro Fuzzy Inference System.” 2017, pp. 1512–1519.

[7] Park, Moon-Hee, et al. “Location-Based Recommendation System Using Bayesian User‟s Preference Model in Mobile Devices.” Ubiquitous Intelligence and Computing Lecture Notes in Computer Science, pp. 1130–1139., doi:10.1007/978-3-540-73549-6_110.

[8] Huang, Yao-Chang, and Shyh-Kang Jenor. “An audio recommendation system based on audio signature description scheme in MPEG-7 Audio.” 2004 IEEE International Conference on Multimedia and Expo (ICME) (IEEE Cat. No.04TH8763), doi:10.1109/icme.2004.139427

[9] S. Kalyania, D., & Ozturk, C. (2011). A novel clustering approach: Artificial Bee Colony (ABC) algorithm. Applied soft computing, 11(1), 652-657.

[10] Kuo, R. J., Syu, Y. J., Chen, Z. Y., & Tien, F. C. (2012). Integration of particle swarm optimization and genetic algorithm for dynamic clustering. Information Sciences, 195, 124-140.

[11]Rana, S., Jasola, S., & Kumar, R. (2013). A boundary restricted adaptive particle swarm optimization for data clustering. International Journal of Machine Learning and Cybernetics, 4(4), 391-400.