International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 06 Issue: 04 | April - 2022 Impact Factor: 7.185 ISSN: 2582-3930

HYBRIDIZATION OF CLUSTERING ALGORITHM FOR BETTER OPTIMIZATION

JAY LIMBACHIYA[1], DARSHAN RAICHADA[2], MANAV SHAH[3], RAJESH BOTHRA[4]

[1] *Student, Computer Engineering, ARMIET College*
[2] *Student, Computer Engineering, ARMIET College*
[3] *Student, Computer Engineering, ARMIET College*
[4] *Professor, Computer Engineering, ARMIET College*

Abstract - In data tunneling and data analyses, clustering plays a vital role. Clustering is a technique wherein the resultant groups formed consists of data which are like one another. Cluster analysis groups testimony based on correlation and diversity among the data elements. These groups formed are called as clusters and it is an unsupervised approach. There are distinct algorithms which consider the attributes of data and the crunch numbers to form clusters from the info. Based on the behavior of the algorithm the centroids are preferred naturally by algorithm or the user can define it. The preferred algorithm for clustering is K-Means which splits up the data based on the mark of compactness, but it also has some demerits like falling in local optimum. So, in order to avoid, that another algorithm which can be used is fuzzy clustering algorithm (FCM). To get hold of fuzzy patterns as a turnout method called fuzzy clustering is utilized. FCM also has another face which describe that the Euclidean distance measures can unevenly weight underlying factors. Getting uplifted from the decorum of birds, particle swarm optimization (PSO) is a worldwide enhancement process. PSO is popularly used in many cluster analysis issues. So, in order to make algorithm gives better results, we are bringing together two algorithms which will take advantage of twain design. This works set forth the hybrid combination of K-means and PSO called as Hybrid-PSO. The motive responsible for linking 2 algorithms is that it gives preferable results in terms of speed and proves to be strong clustering algorithms which will be shell out fitter optimization.

Key Words: Fuzzy c-means; K-Means Clustering algorithm; Particle swarm optimization.

1.INTRODUCTION

A supervised learning algorithm which groups the info into predefined classes is called as classification. Clustering is a way of unsupervised learning algorithm which tries to group the data together were in the resultants groups (clusters) formed has resemblance between them. When categorization of data does not rely on any predefined class, then it is said to be unsupervised type of data. K-Means is a famous clustering analysis algorithm which was invented in the year 1965. It can handle large number of datasets so in the field of data mining it is used more often. It works randomly on the selection of clusters which are called centroids. The original method tries to merge the local minimum. Madhu Yedla [1] has urged a scheme where they have did one's best to improve the original style in terms of veracity and also to find way to appoint correct initial centroids. A newer way of allocating data points to all the sufficient clusters which are within less time spam is also told.

There are many clustering algorithms to find quick fix of clustering dispute. In order to obtain productive and severity as profit in large datasets, k-means is the best clustering algorithm given by Navjot Kaur [2]. Using ranking based method, they have tried to improve the performance of k-means algorithm. To get fuzzy patterns from data as output fuzzy clustering method is applied. Application in the field of image analysis. Even though FCM is successful, but we should keep in mind of few issues that must be deal with in practical utilization of these algorithms.

There are various types of data available and many application areas of the same in which clustering technique can be applied. PSO algorithm came into account in the year 1995 and is easy to implement. PSO is inspired from the behavior of birds or fish school. Here the velocity of every particle keeps on changing gradually so it has a dynamic behavior. Calculation in PSO is easier and it is also skilled in global investigation. PSO finds the molecule best position (pbest) found up until this point and the best situation inside the area of that molecule (gbest). Every molecule stores its present area, current speeding up and its best position found up until this point and afterward further chips away at it.

The remainder of the study is cataloged within the following fashion of section 2 has literature survey, section 3 consists of existing system, section 4 has the proposed system followed by conclusion.

2. LITERATURE SURVEY

Ka-Chun Wong [3] has made a brief study on data clustering algorithms where right from the design concept to approach, different clustering prototype are discussed. Various characteristics of methods are classified, and the performance metrics of clustering is also mentioned. The need for clustering and the application areas where it could be applied are also stated. The urge to gain better results in the application areas where clustering is applied is the main objective. By utilizing clustering algorithms, the works gets easy and fast. Chintan Shah and Anjali Jivani [4] have used data mining algorithms to predict forest fire. They have organized comparison based on WEKA (The Waikato Environment for Knowledge Analysis). They have correlated

clustering algorithms based on partition method, hierarchy method and density method. The uncomplicated version for partitioning is K-Means. It works on predefined classes, and its drawback is that there is no exact explanation to catch least number of clusters on inclined dataset. Hierarchy is used because it put in order the objects in a tree like structure and follows bottom-up approach. AGNES (Agglomerative Nesting) is used to predict forest fire. To identify the random shaped clusters, density-based method is used, and it works well on any modest or average level datasets and showed K-Means is the best suitable in their work.

Tao Lie et.al [5] have granted FCM algorithm in a distinct way which is better than its own prior version. As known FCM is tested in image segmentation to truncate the influence of noise on image, here they have used FR-FCM which will recuperate the segmentation process .An upgraded FCM finding situated on morphological reconstruction and membership draining is more rapid and booming than FCM. They have demonstrated that this coming is more suitable than either of improved FCM algorithm.

Weijia Lu [6] has tried to work on comprehensive datasets under Hadoop skeleton using K-Means clustering algorithms. The prime focal point was to refine the skill and productivity but on enormous datasets. A cumulative miniature of K-Mean is applied based on its density. As proposed, they are working on bulk of data, the algorithm segregates the whole block of info into parts of clustering blocks using weighted K-mean and fusion K-mean. But it is a time exhausting approach. Instead of this drawback their results show that the accuracy of algorithm is surpassed by more than 10% than other two clustering algorithms. A hybrid clustering path based on K-Mean and Ant Lion Optimizer is proposed for optimal cluster analysis. ALO is a global optimization model. Santosh Kumar [7] has recycled a k-mean algorithm with ALO and has compared the results of proposed method with various datasets and has concluded that the designed model gives finer output than rest of the algorithms mentioned. K-Means is a well-known group investigation technique which expects to parcel various information focuses into K bunches. It has been effectively applied to various issues. Be that as it may, the effectiveness of K-Means relies upon its in statement of group focuses. Diverse multitude insight methods are applied to grouping issue for improving the exhibition. In this work a half and half bunching approach dependent on K-means and Ant Lion Optimization has been advised for ideal group investigation. Subterranean insect Lion Optimization (ALO) is a stochastic worldwide streamlining image. The presentation of the urged calculation is analyzed against the exhibition of K-Means, K-Means-PSO, K-Means-FA, DBSCAN and Revised DBSCAN bunching techniques dependent on various execution measurements. Experimentation is performed on eight datasets, for which the measurable investigation is done. The acquired outcomes show that the half and half of K-Means and Ant Lion Optimization strategy operate ideally superior to the other three calculations as far as entirety of intra-cluster separations and F-measure. Also, they have given a statement based on results, that the proposed algorithm in terms of accuracy achieves 90%.

Chaoyang Zhang [8] says traditional grouping calculations utilize all the clue to gain proficiency with the bunch arrangements. Be that as it may, in genuine world applications, a minor data represents cognizant conduct and can be summed up, whereas some other information present powerless propensities to be appointed to a definite example. For such case, this work presents knowledge determination system for K-Means calculation to get more veracity bunches through information assortment. It varies in three regards from customary k-implies type computation. Initially, in the bunch learning process, we take the revised estimation of Bregman Information of group, which is created by blending one information into the possible groups, as the proportion of information thing's bunching propensity. Second, just information things with solid bunching propensities, that is the changed estimation of group's Bregman Information is not exactly the predefined range, are chosen to get recognizable with the group designs, while the other information of focal point are overlooked and have a place with no group.

Saptarshi Sengupta [9] says fluffy bunching has become a broadly utilized information mining procedure and assumes a significant job in gathering, navigating and specifically utilizing data for client determined operations. The deterministic Fuzzy c-means calculation may appear in problematic arrangements meanwhile enforced to multidimensional information in genuine world, time-obliged issues. In this paper the Quantum-carried on Particle Swarm Optimization (QPSO) with a completely associated topology is combined with the Fuzzy C-Means Clustering calculation and is tried on a set-up of datasets from the UCI Machine Learning Repository. The worldwide inquiry capacity of the QPSO calculation helps in evading stagnation in nearby optima while the delicate grouping approach of FCM assists with parceling information dependent on participation probabilities. Bunching execution records, for example, F-Measure, Accuracy, Quantization Error, Inter-cluster and Intra-cluster separations are accounted for serious procedures, for example, PSO K-Means, QPSO K-Means and QPSO FCM over all datasets considered. Trial results show that QPSO FCM gives similar and as a rule better outcome when thought about than the others

2.1 EXISTING SYSTEM

The k-means algorithm can be considered as a hard-clustering algorithm and it is the most famous partitioning clustering algorithm. The k-means model is considered to be not suitable for the data sets where there is no clear-cut borderline in the middle of the clusters [10].

Advantages of K-means:

- K-mean clustering is straightforward as well as versatile.
- K-mean clustering act straightforward.
- For the variables with large value, if we keep k smaller then K-Means can outperform hierarchical clustering concerning speed.

Disadvantages of K-means:

- Forecasting K-Value is challenging.
- It does not function properly, with the comprehensive bundle.
- Based on the value of K, the results of clustering are dependent [11].

Traditional FCM acts as a crucial clustering algorithm. The iterative process gets trapped into the local optima due to the aimless collection in center points.

Advantages of FCM:

- This algorithm is finer than the k-means algorithm.
- This algorithm works well with overlay info.

Disadvantages of FCM:

- When the value of β is less the results are good, but the number of iterations is more.
- Euclidean distance measures can unequally weight underlying factors.

Taking motivation from the performance of birds, Particle swarm optimization (PSO) may be a population-based speculative optimization finding. PSO is vital algorithm in the grade of Swarm Intelligence (SI), where partnership and contact between the bird communities, allows the population to imitate the general markings of society. In clustering analysis PSO is applied [12].

Advantages of PSO:

- In the fields of research and engineering, PSO is exercised.
- PSO does not change and do any projecting calculation. With the help of the speed of particle the search takes place.
- PSO accepts the important number code, which is set directly by the answer.
- Calculation in PSO is easier and skilled in global investigation.

Disadvantages of PSO:

- It gradually merges in the polished search stage.
- It has infirmed local search skills.
- The method has a disadvantage that it cannot do its job on the affairs of non-coordinate.

2.2 PROPOSED SYSTEM

The conventional K-mean algorithm calculation depends on disintegration, most generally utilized in information mining field. The idea is use K as a boundary, divide n object into K groups, to make generally high closeness in the bunch, moderately low likeness between groups. Also, limit the complete separation between the qualities in each bunch to the group place. The group focus of each bunch is mean estimation of group. The count of closeness is finished by mean estimation of the bunch objects [13]. The estimation of the comparability for the calculation choice is by the complementary of Euclidean separation. In other words, the closer the separation, the greater the similitude of two articles, and tight clamp versa. The calculation comprises of two separate stages. The main stage chooses k focuses arbitrarily, where the worth k is fixed ahead of time. The following stage is to take every information article to the closest focus. Euclidean separation is commonly considered to decide the separation between every information object and the group places. At the point when all the information objects are remembered for certain bunches, the initial step is finished, and an early gathering is finished. Recalculating the normal of the early framed groups. This iterative procedure proceeds over and standard again until the capacity turns into the base.

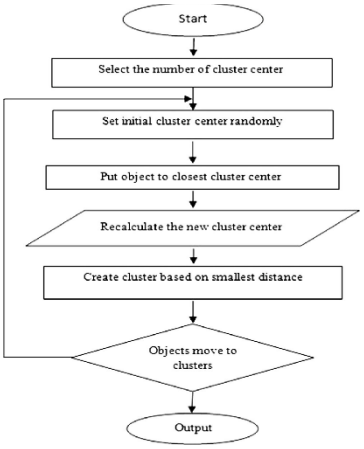

Fig1: Flow chart of K-mean algorithm

PSO is a speculative approach having a chance of converging at an early stage and even with sub-optimal solutions. If we must get a good solution, then the PSO-based clustering algorithm needs training coefficient tuning. Moving back to the context of clustering, the answer can be defined as a group of coordinates such that all correspond to a cluster centroid concerning the c-dimensional position. There is a chance of one possible solution for the PSO-Clustering problem. In that c-dimensional cluster centroid can be observed in every n solution. Even though the algorithm can be utilized in any dimensional space, only two-dimensional or three-dimensional spaces are used for visualization. Our proposed algorithm aims at evaluating the given fitness function in the simplest possible way. Precisely for our context, it should try finding the modest spatial configuration of centroids. As an edge is represented by each particle using the n-dimensional space, the expectation should be to keep the consistency of its position with the favorite area of the particle and within the region of that fragment. There are 2 methods to calculate the similarity here we are using Euclidean distance.

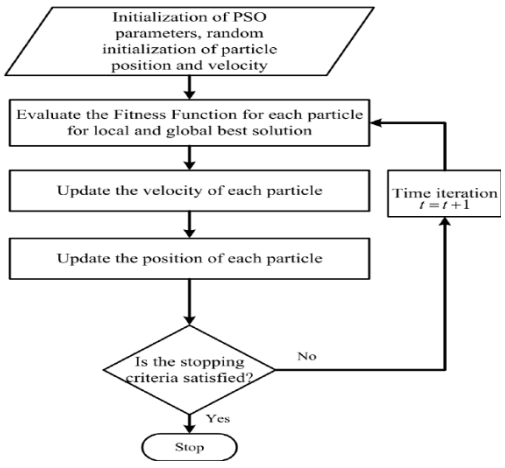

Fig2: Flow chart of PSO algorithm

This can be only possible when the values stored by each particle are x_i: current location, v_i: current acceleration and yi: best position obtained so far [14].The expression for the adjustment of the position of the particle can be given as follows.

$$v_{i,k}\,(t+1) = w v_{i,k}\,(t) + c_1 r_{1,k}\,(t)(y_{i,k}\,(t) - x_{i,k}\,(t))$$

$$+ c_2 r_{2,k}(t)(\hat{y}_k(t) - x_{i,k}(t)) \quad (1)$$

$$x_{i,k}(t+1) = x_{i,k}(t) + v_{i,k}(t+1) \quad (2)$$

In this expression, inertia weight is given by w, acceleration constants are represented by c1 and c2. Similarly, the samples from the Uniform Distribution are given by $r_{1,k}(t)$ and $r_{2,k}(t)$.

The personal best location of fleck, defined to be the area which gives the utmost evaluation of the fitness function over all instances, is modernized as:

$$y(t+1) = \begin{matrix} y_i(t) \text{ if } f(x_i(t+1)) \geq f(y_i(t)) \\ x_i(t+1) \text{ if } f(x_i(t+1)) < f(y_i(t)) \end{matrix} \quad (3)$$

The fitness function is defined in terms of the quantization error as shown in this equation no (4). Each particle is constructed as: $x_i = (m_{i1},\ldots,m_{ij},\ldots m_{iN_C})$ where m_{ij} is the cluster centroid of the i-th particle in cluster C_{ij}.

$$J_e = \frac{\sum_{j=1}^{N_c} \left[\sum_{z_p \in C_{ij}} d(z_p, m_j)/|C_{ij}| \right]}{N_c}$$

The PSO is typically executed within endless repetitions of the Equation 1 and Equation 2, until a specified number of repetitions has been reached. An alternate solution is to prevent when the velocities are on the brink of zero, which suggest that the algorithm has reached a least possible within the optimization process. Another time, it's important to note that albeit in 1two kinds of PSO approaches are presented, respectively named g-best and l-best where the social components is essentially bounded either to the present neighborhood of the particle instead to the whole swarm, during this work we refer only to the essential g-best proposal. Dynamic clustering is also one the approach stated by Thushara, K wherein they have used zoned based approach [15].

So, our work presents the consolidation of K-mean and PSO algorithm, which will manage the merits of both algorithms called as Hybrid-PSO. In this hybrid algorithm, K-mean is carried out first and the output gained from this is feed to one of the specks in PSO clustering algorithm and then the PSO algorithm is carried out.To solve the fuzzy clustering problem HYBRID-PSO algorithm is as follows:

Hybrid-PSO for fuzzy clustering:

Step 1. Number of specks = 10

Step 2. Deal with K-means on the data and grant the calculated centroid to one particle.

Step 3. Load other nine speck to have randomly selected N_c *cluster* centroids

For *i* in range t_{max}:

a) For j in range No. of fleck:

i. Individual vector knowledge.

A) Count the Euclidean distance d (z_p, m_{ij}) to all cluster centroids C_{ij}

B) Appoint the data vector to the cluster such

ii. Count the fitness function using equation 4.

b) Amend local best position using equation 3

c) Amend the global best position as the position of speck which decreases the fitness function

d) Renew the cluster centroids using equation 1, 2.

As this hybrid algorithm is combination of individual existing algorithm, so the mathematical formulation of hybrid would be the combination of similar mathematical formulation described in the existing system.

We have applied this algorithm to test on wine database. This is the data-set that is acquired as the aftereffects of a substance examination of wines developed in the locales of Italy yet got from three unique areas. The investigation decided the amounts of 13 constituents (inputs) found in every one of the three kinds of wines (1, 2, 3, 4) (7) (8). The wine dataset contains the aftereffects of a synthetic investigation of wines developed in a region of Italy. Three kinds of wine are spoken to in the 178 examples, with the consequences of 13 concoction investigations recorded for each example. The Type variable has been changed into a category variable. The information contains no missing qualities and consists of just numeric information, with a three-class target variable (Type) for order.

By using this hybrid approach, we have overcome the drawback of K-Means with improved error rate and good convergence speed. And also it requires less iteration.

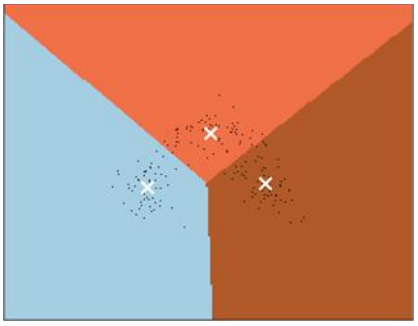

Fig3: K- Mean clustering on wine dataset.

The figure3 shows results of clustering using k-means algorithm and the centroids are noticeable with white cross.

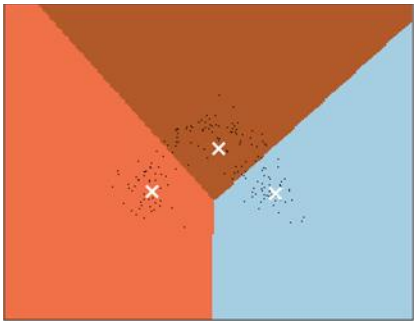

Fig4: PSO clustering on wine dataset.

The Figure4 shows the result of clustering using PSO algorithm, here also centroids are noted with white cross.

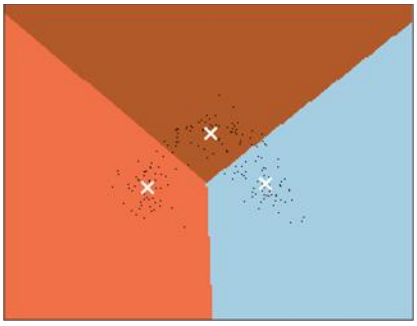

Fig5: Hybrid clustering on wine dataset.

Figure5 shows the result of hybrid algorithm (k-mean - pso). Where centroids are marked with white cross.

The table 1 below shows the results obtained when we executed the Hybrid-PSO algorithm explained above, it calculates the silhouette score for all the three algorithms and the number of iterations performed is 1000. As the number of iterations will expand the outcome will get preferable.

TABLE I: Algorithm Comparison

Algorithm	Quantization Error	Silhouette Score
K-means	0.498	0.300
PSO	0.718	0.036
Hybrid-Pso	0.498	0.301

The silhouette score ranges from +1 to -1. Where in the positive score means it is better and the negative score means it is not good. Character near 0 indicate imbricated clusters. Weak values generally pinpoint that a sample has been deputize to the false bundle, as a contrasting chunk is more related.

3. CONCLUSIONS

The fuzzy c-means algorithm is very delicate to initialization and gets easily captured in local optima. K-mean system yields more accurate and reproducible results compared to FCM. K-mean methods takes a bit more time to execute because it provides better and accurate results than FCM. On the other side, the particle swarm method is a global stochastic design that can be invoked and adjusted calmly to iron out various function optimization dilemma.In this work, a new usage for hybridization of K-mean and PSO for using in clustering analysis is presented. The achievement of Hybrid-PSO will be set side by side with the original PSO clustering and k-mean algorithm in charge of exactness and rapidness. Hence, we can conclude that the proposed idea results in a good clustering algorithm and the hybrid combination gives better optimization than k-mean and PSO in terms of error rate. As the number of iterations increases the results will be more accurate only the point is that it will be a bit time consuming task, but the results will be preferable. Here the proposed algorithm also depends on the type of dataset you are using, so when this algorithm is applied on some other datasets the results may vary too.

ACKNOWLEDGEMENT

We are grateful to all those who helped us make this paper, for the valuable advice provided by them in their respective fields. We are grateful for their co-operation during the scripting of the paper.

REFERENCES

[1] Madhu Yedla, Srinivasa Rao Pathakota, T M Srinivasa, "Enhancing K-means Clustering Algorithm with Improved Initial Center", International Journal of Computer Science and Information Technologies (IJCSIT), Vol. 1 (2), 2010, 121-125.

[2] Navjot Kaur, Jaspreet Kaur, "Efficient k-means clustering algorithm using ranking method in data mining", International Journal of Advanced Research in Computer Engineering and Technology, volume 1, Issue 3, may2012.

[3] Ka-Chun Wong, "A short survey on data clustering algorithms", Second International Conference on Soft Computing and Machine Intelligence, November 2015.

[4] Chintan Shah, Anjali Jivani, "Comparison of data mining clustering algorithms", Nirma University International Conference on Engineering 2013.

[5] Tao Lie, Xiaohong Jia, Yanning Zhang, Lifeng He, "Significantly fast and robust fuzzy c-means clustering algorithm based on morphological reconstruction and membership filtering", IEEE Transactions on Fuzzy System 2018.

[6] Weijia Lu, "Improved k-means clustering algorithm for big data mining under Hadoop parallel framework", Journal of Grid Computing December 2019.

[7] Santosh Kumar,Shubhra Biswal Majhi, S.K.,Biswal, "Optimal cluster analysis using hybrid k-means and ant lion optimizer", International Journal of Modern Science 2018.

[8] Chaoyang Zhang, "Fuzzy systems and knowledge discovery", 13th International Conference on Natural Computation 2017.

[9] Saptarshi Sengupta, Sanchita Basak, Richard Alan Peters II, "Data clustering using a hybrid of fuzzy c-means and quantum-behaved particle swarm optimization", IEEE 8TH Annual Computing and Communication Workshop and Conference (CCWC) 2018.

[10] Maryam Lakshkari, Mohammad Hossein Moattar, "The improved k-means clustering algorithm using the proposed extended pso algorithm", International Congress on Technology, Communication and Knowledge (ICTCK), November 2015.

[11] Ali Maroosi, Somayeh Geravand, "A novel threshold-based clustering method to solve k-means weaknesses", International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), August 2017 IEEE.

[12] Vicente Rueda, Maria Alducin, Efren Montes, Nicandro Ramirez, "Particle swarm optimization with feasibility rules in constrained numerical optimization-a brief", International Autumn Meeting

on Power, Electronics and Computing (ROPEC), November 2016.

[13] Shi Na, Liu Xumin, Guan Yong, "Clustering algorithm: an improved k-mean clustering algorithm", Third International Symposium on Intelligent Information Technology and Security Informatics, April 2010 IEEE.

[14] Qian Zhang, Xing Li, Quang Tran, "A modified particle swarm optimization algorithm", International Workshop on Education Technology and Training, August 2005.

[15] K. Thushara, Jennifer S. Raj, "Dynamic clustering and prioritization in vehicular ad-hoc networks:zone based approach", International Journal of Innovation and Applied Studies, vol.3, no.2, pp. 535-540, june 2013.