International Journal of Research in Engineering, Science and Management
Volume-3, Issue-9, September-2020
journals.resaim.com/ijresm | ISSN (Online): 2581-5792 | RESAIM Publishing

Enhance Clustering Algorithm Using Optimization

Roshankumar Ramashish Maurya[1*], Anand Khandare[2]
[1]*Department of Computer Engineering, Thakur College of Engineering and Technology, Mumbai, India*
[2]*Associate Professor, Department of Computer Engineering, Thakur College of Engineering and Technology, Mumbai, India*
*Corresponding author: roshan.maurya90@gmail.com

***Abstract*: Unsupervised learning can reveal the structure of datasets without being concerned with any labels, K-means clustering is one such method. Traditionally the initial clusters have been selected randomly, with the idea that the algorithm will generate better clusters. However, studies have shown there are methods to improve this initial clustering as well as the K-means process. This paper examines these results on different types of datasets to study if these results hold for all types of data. Another method that is used for unsupervised clustering is the algorithm based on Particle Swarm Optimization. For the second part this paper studies the classic K-means based algorithm and a Hybrid K-means algorithm which uses PSO to improve the results from K-means. The hybrid K-means algorithms are compared to the standard K-means clustering on two benchmark classification problems. In this project we used Kaggle dataset to with different size (small, large and medium) for comparison PSO, k-means and k-means hybrid.**

***Keywords*: Clustering, K-means clustering, Particle Swarm Clustering.**

1. Introduction

In unsupervised learning, training methods do not use any kind of labels between algorithms. This can reduce the time required to differentiate training, and allows researchers to see the properties in the data. One of the methods for unsupervised learning is the K-means method, which divides data into separate group's k. Each collection is thought to be Gaussian and circular, with each data point in the collection closest to its center.

The traditional method for initializing the K-means method is to randomly assign cluster centers and let the algorithm distribute those random centers to appropriate locations. However, depending on the data structure this does not always create predictable clusters after training. A refined initialization method has been developed by Bradley and Fayyad that refines the random initial clusters.

Refined collections are used in the K-means algorithm to separate data. The first refined collections are designed to produce unpredictable collections. Particle Swarm Optimization based clustering algorithm was used for the integration of vector image and data. This paper will compare the hybrid K-means algorithm against the standard PSO and the K-means standard (scikit package) algorithm, in addition to heart disease, breast cancer, diabetes, wine quantity and MNIST information data and attempts to use a different type of the database.

2. Background and Motivation

A. *Background*

Integration is one of the most challenging methods of mining in the data acquisition process. Managing large amounts of data is a daunting task because the goal is to find the right subdivisions in an unsupervised manner (i.e. without prior knowledge) in an attempt to maximize internal cluster similarities and reduce cluster similarities that also keep high cluster mergers. Data collection takes place in subsets in such a way that similar conditions are collected together, while different conditions belong to different groups.

Circumstances are thus organized into a more effective presentation that reflects the number of samples. Therefore, the release of cluster analysis is the number of groups or clusters forming the composition of the division, of the data set. In short collections the data processing process has become a sound mathematical analysis group. Exploitation of Data Mining and Knowledge access has permeated a variety of Machine Learning System.

B. *Motivation*

As the number of digital documents over the years as the Internet has grown exponentially, managing information search, and retrieval, etc., have become more important issues. Advanced methods of organizing large volumes of random text into small numbers of logical clusters will be of great help to combine such as indexing, filtering, default metadata production, number of web resource catalogues and, generally, any program that requires text editing.

There are also a large number of people who are interested in reading certain stories so there is a need to compile news articles from the number of available articles, because a large number of articles were added to each data and many articles were related to the same issues but included in different sources. By compiling articles, we can narrow down our search domain with recommendations as most users are interested in issues related to a few groups.

This can improve the effect of time efficiency on a large scale and can help identify similar issues from different sources. The main motivation is to compare the different types of unattended algorithm to learn how they behave, their advantages, and their disadvantages and to learn how to choose an unattended learning algorithm depending on the type of dataset. This paper projected we describe our hybrid K-means clustering algorithm flow, compare and analysis their behavior on two types of dataset.

Also implement the different parameter of unsupervised learning algorithm to observed error rate, Silhouette score, by compare Hybrid K-means clustering algorithm with standard PSO algorithm and K-means algorithm we get their advantage and disadvantage.

3. Method Description

A. Standard K-means

The modified startup method uses a set of categories J of data. Each of these sub-sections is designed to randomly select a small percentage of the original data. From each clause is obtained a set of k-center centers, and any empty collections are given a point with a great distortion and then reassembled the whole clause. When all subdivisions have empty collection centers, the *J * k* points are grouped using random startups.

The result of this integration is used as the first K-means collection centers throughout the database. The first was the purity of the middle class, a measure based on data labels. The second was a distortion, or a double L2 range of data, of groups where the L2 / Euclidean range is given as:

$$\mathrm{d}(\mathbf{p},\mathbf{q}) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2} \qquad (1)$$

Flow Chart:

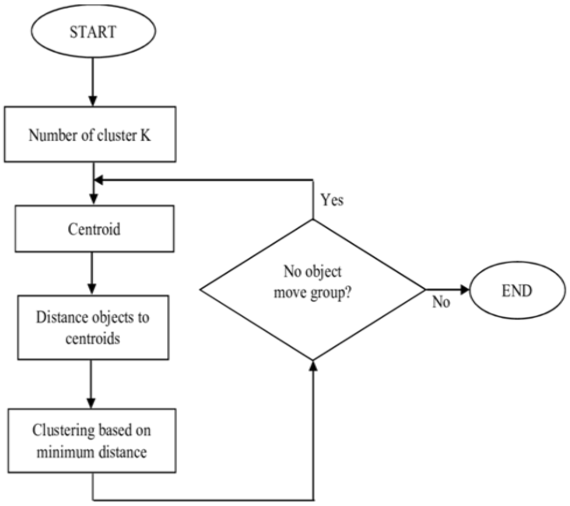
START
Number of cluster K
Centroid
Distance objects to centroids
Clustering based on minimum distance
No object move group?
Yes
No
END

Fig. 1. K-means data flow

This paper uses the silhouette score as a measure of quality. This score, from -1 to 1, compares the inter-cluster distance of data to the distance to the nearest cluster. A negative score represents mis-clustered data, with points assigned to a cluster that should be in another. A positive score represents defined clusters, with a higher score meaning more distinct clusters. A score of 0 represents overlapping clusters.

To truly investigate the difference between the random and refined initialization, and to compare PSO algorithms, K-means algorithm with hybrid K-means, 5 different types of datasets were compared. From the UCI Machine Learning Repository, the Heart Dieses Dataset, Breast Cancer Diagnostic, Diabetes, Wine Quality, MNIST datasets were used.

- *Heart dieses dataset:* This database contains 76 attributes, but all published experiments refer to using a subset of 14 of them. This dataset we use with 300 rows and 14 columns.
- *Breast cancer diagnostic dataset:* The Database are: This is numeric based dataset. In this dataset define the tumour and cancer score. This dataset has 600 rows and 14 columns.
- *Diabetes dataset:* Diabetes is a sparse dataset with 2 classes, 10000 features and 900 points which examine the effect of number of features used.
- *Wine quantity dataset:* This dataset is well studied and "well-behaved". It has 13 features, 3 classes and 178 samples. Hence a good problem for studying the differences in initialization and comparison of algorithms.
- *MNIST dataset:* This dataset too has been studied extensively and has well documented behaviour. It has 16 features, 10 classes and 1797 samples (10992 points).

B. Standard Particle Swarm Optimization

The PSO was inspired by the social behavior of bird populations and was first developed by Eberhart and Kennedy in 1995 It is a man-made foundation where the algorithm stores particles each representing a solution to the problem of efficiency. The PSO aims to obtain a particle position that provides an excellent test of a given performance function. The next section describes the performance of Particle Swarm Optimization and surpasses the integration of PSO and K-mean PSO collection algorithms. For this purpose, the following symptoms are described:

N_d : Dimension of data vector
N_c : Number of cluster centroids
z_p : p^{th} Data vector
m_j : Centroid vector of cluster j
C_j : Subset of data vector that form cluster j

One of the key features of the compilation is the similarity scale used to combine the data with the number of pre-determined collections. Two outstanding methods used to install a computer are similar to the Euclidean range, which is used for data vector integration, and the cosine aggregation process, which is used for document integration. The Euclidean range is used as a measure of similarity. The data vectors within the collection are in a small 'Euclidean' range from each other, and are associated with one centroid vector of that collection. The vector distance in centroid is determined using equation 1:

$$d(z_p, m_j) = \sqrt{\sum_{k=1}^{N_d} (z_{pk} - m_{jk})^2} \qquad (2)$$

Flow Chart:

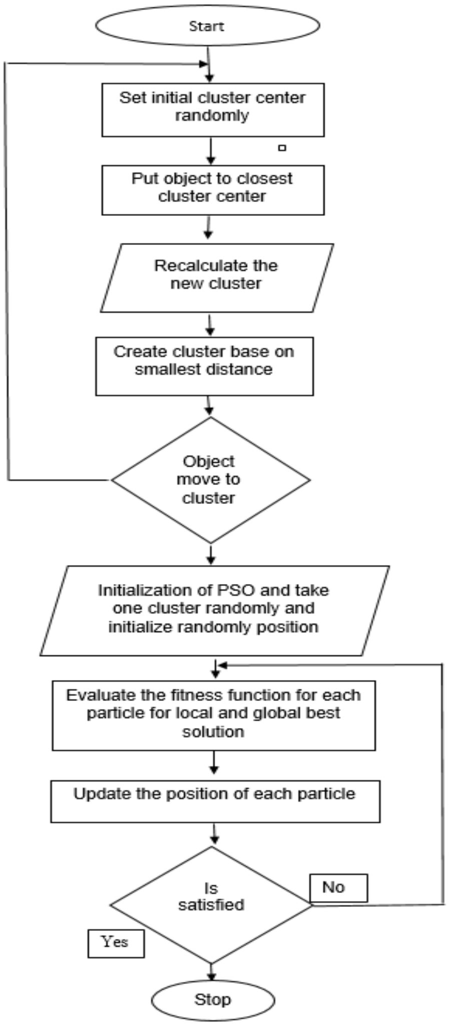
Start
Set initial cluster center randomly
Put object to closest cluster center
Recalculate the new cluster
Create cluster base on smallest distance
Object move to cluster
Initialization of PSO and take one cluster randomly and initialize randomly position
Evaluate the fitness function for each particle for local and global best solution
Update the position of each particle
Is satisfied
No
Yes
Stop

Fig. 2. Hybrid K-means data flow

Algorithm initially start with a set of randomly generated points where each point refers to the position of a particle in N_d dimensional space. Associated with each particle is its velocity vector. Each particle has the following information x_i : The current position of the particle v_i : The current velocity of the particle; y_i : The personal best position of the particle. A particle’s position at the next time instance is then calculated as:

$$v_{i,k}(t+1) = wv_{i,k}(t) + c_1 r_{1,k}(t)(y_{i,k}(t) - x_{i,k}(t)) + c_2 r_{2,k}(t)(\hat{y}_k(t) - x_{i,k}(t)) \qquad (3)$$

$$x_{i,k}(t+1) = x_{i,k}(t) + v_{i,k}(t+1) \qquad (4)$$

Where, w is the inertia weight, c_1 and c_2 are the acceleration constants,r_1, j(t), r_2, j(t) ~ U(0, 1) and k = 0,..., N_d. As is clear from equation 3, the velocity is updated based on three components: first is a fraction of its previous velocity, second is cognitive component which is a function of the distance of particle from its personal best position and third is social component which is a function of distance of particle from the global best position. The personal best position of a particle, defined to be the position which gives the best evaluation of the fitness function over all instances, is updated as:

$$y_i(t+1) = \begin{cases} y_i(t) \text{ if } f(x_i(t+1)) \geq f(y_i(t)) \\ x_i(t+1) \text{ if } f(x_i(t+1)) < f(y_i(t)) \end{cases} \qquad (5)$$

C. Hybrid K-means Clustering

Hybrid K-means algorithm is a hybrid of K-means and PSO methods of clustering. In this, K-means is executed once and the results of K-means are used to seed one of the particles in PSO clustering algorithm. Then PSO algorithm is executed.

Algorithm for Hybrid K-means clustering:

```
1) Number of particles = 10
2) Execute K-means on the data and assign the calculated
   Centroid to one particle
3) Initialize other nine particles to have randomly selected
   $N_c$ cluster centroids.
4) For i in range $t_{max}$ :
 a) For j in range No. of particles:
    i) For each data vector:
          A. Calculate the euclidean distance
 $d( z_p,m_{ij})$ to all cluster centroids $C_{ij}$.
          B. Assign the data vector to the cluster such
             that the euclidean distance is minimum.
    ii) Calculate the fitness function.
 b) Update local best position using equation 5.
 c) Update the global best position as the position of
 particle which minimizes the fitness function.
 d) Update the cluster centroids using equation 3, 4.
```

4. Experimental Results

A. Hybrid K-means vs Standard K-means vs PSO comparison

In this section we will discuss the effects of silhouette symbols from each information. In each test, the random K-means method used was from the Scikit-learn package of Python. This was also the basis for a random launch within a fixed path. Each collection is limited to 50 iterations. Adjusted data subsets are 10% of each of the original data. The Heart Dieses database was tested with a range of 300 lines within a refined database and this hybrid k-methods method provides an additional 10% accuracy.

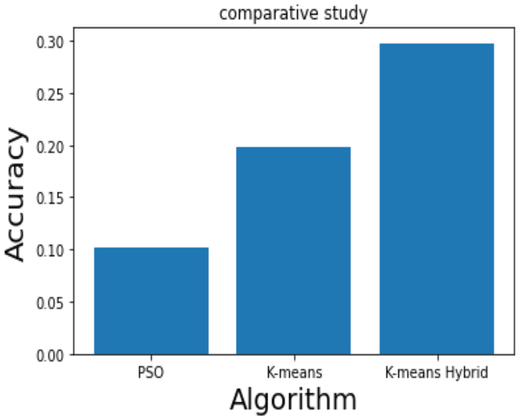

Fig. 3. Heart dieses dataset accuracy

The number of collections also varies. The results show that the 2 methods are comparable in their peaks. This dataset is very complex, as it has a limited number of both features and categories. This is a lack of complexity that can cause similar peaks. Each method is able to identify three different classes, similar to true symbols. However, when the number of collections did not equal the number of classes.

The Breast Cancer Diagnostic Dataset we used to 600 rows large dataset to compare this 3 dataset and again we get 10% more Silhouette score at 3 clusters and less error rate.

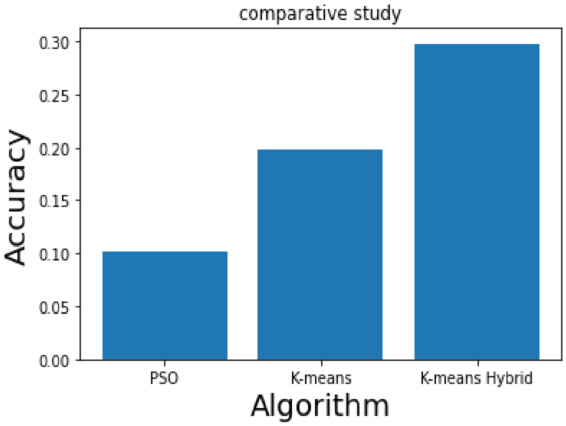

Fig. 4. Breast cancer dataset accuracy

The next dataset we used for comparison its Diabetes datasets with 800 rows.

The same method of comparison as was used on the MNIST and Wine Quantity dataset was used, with the exception of extending the range of clustering's. Since the MNIST set has 10 classes, the range of tested clusters needed to be larger. The number of class labels the set has does not directly affect the algorithm, since these are unsupervised learning methods the training of the clusters does not use the labels. However, the structure of the data is more complex, with a comparable number of features as the Wine Quantity dataset. The results of this set are shown in figure 6 and 7.

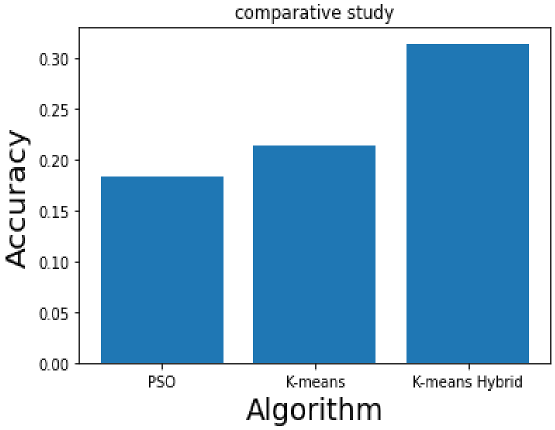

Fig. 5. Diabetes dataset accuracy

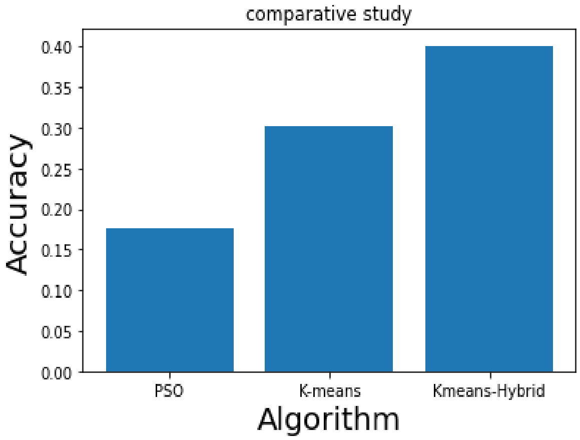

Fig. 6. Wine quality dataset Accuracy

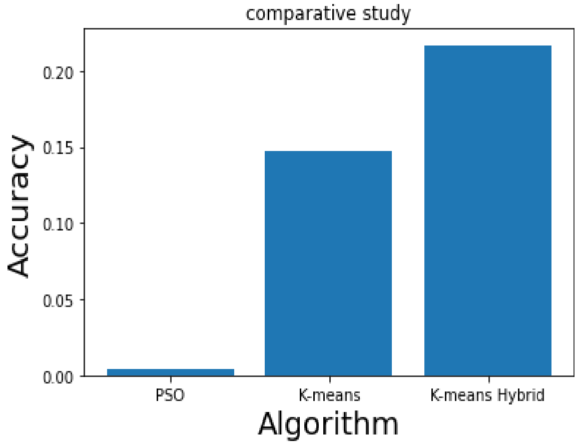

Fig. 7. MNIST dataset accuracy

This setting indicates why a modified start up method may be preferred to a random process. While there are many calculations at the beginning of training, it can produce better performance. This extended functionality will indicate if the database is complex enough, while simple data sets may not see any improvement at all between the two approaches.

All five refined databases will work and even better than the standard k-and PSO methods given a fair amount. The mathematical method of scoring uses the concept of distance. As the features are removed, the size of the feature space is removed causing smaller distances between points. However, this will affect random and equally refined methods, allowing for comparisons between them and hybrid k-means give better result as compare to standard k-means and PSO clustering algorithm.

This very reduced feature space makes the structure simplified, which has resulted in results. The gap between performance measures is very large from 8 to 20 factors. This region shows a refined approach that can maintain a better definition of the collection as the difficulty increases. There is a satisfying point where the methods work the same way.

Table 1
Algorithm Comparison

Dataset	Algorithm	Error rate	Silhouette score	elapsed time
Heart Dieses Dataset (300 rows)	PSO	1.4706	0.1022	0.0140
	K-means	1.1882	0.1976	0.0305
	Hybrid K-means	1.1751	0.2976	0.3027
Breast Cancer Dataset (600 rows)	PSO	1.4815	0.3062	0.0153
	K-means	1.0601	0.3375	0.0511
	Hybrid K-means	1.0585	0.4375	0.1045
Diabetes Dataset(800 rows)	PSO	0.9836	0.1832	0.0192
	K-means	0.8647	0.2139	0.0479
	Hybrid K-means	0.7588	0.3139	0.0626
Wine Quality Dataset (1000 rows)	PSO	0.7301	0.1757	0.0123
	K-means	0.4982	0.3013	0.0831
	Hybrid K-means	0.4878	0.4013	0.2112
MINIST Dataset (3000)	PSO	36.0276	0.0044	1.0889
	K-means	97.7099	0.1471	0.8159
	Hybrid K-means	25.3278	0.2169	1.9922

Table 1 lists the performance of three algorithms in the Wine and Digital databases limited over 10 simulations. One thing to note here is that although the Quantization error can be compared to the given database algorithms, it is not comparable between different databases. This is because quantization error depends on the number of clusters, the pre-processing of data, the number of samples among other things that are very different from the data sets.

From the wine database it is clear that the PSO performs worse than the K methods when compared to the quantization error and silhouette score. However significant improvements can be seen in the Hybrid K-means algorithm. When a single particle in a PSO algorithm is sown with results from the K-means algorithm, the resulting algorithm works much better than the original PSO and is much better than the standard random K-means.

5. Output

A. Output for heart dieses dataset

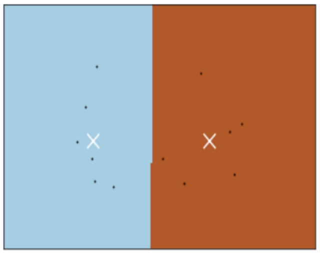

Fig. 8. K-means Clustering

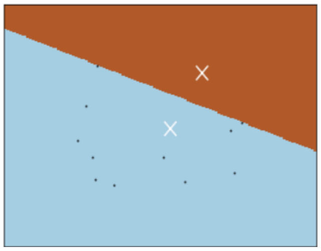

Fig. 9. PSO Clustering

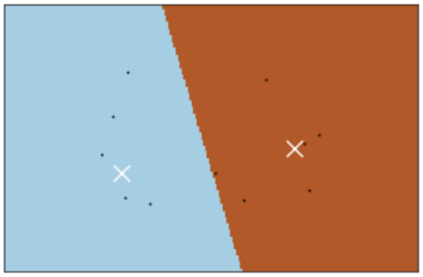

Fig. 10. Hybrid K- Mean Clustering

B. Output for breast cancer diagnostic dataset

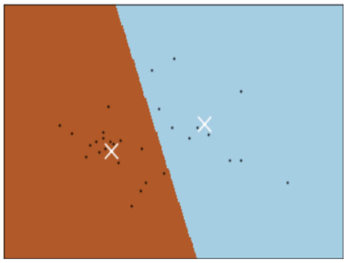

Fig. 11. K-means Clustering

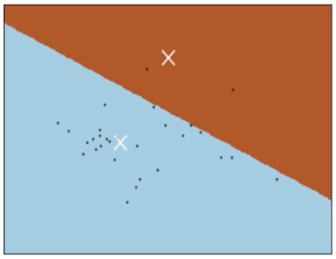

Fig. 12. PSO Clustering

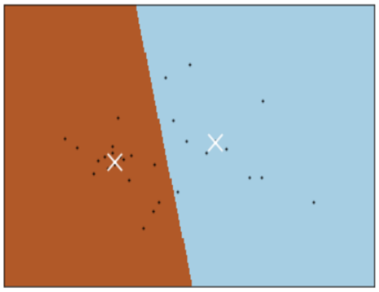

Fig. 13. Hybrid K- Mean Clustering

C. *Output for Diabetes Dataset*

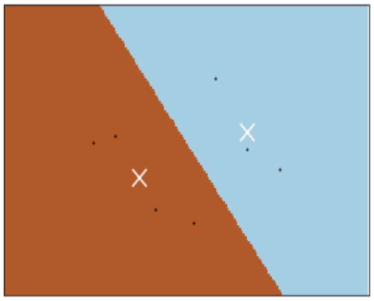

Fig. 14. K-means Clustering

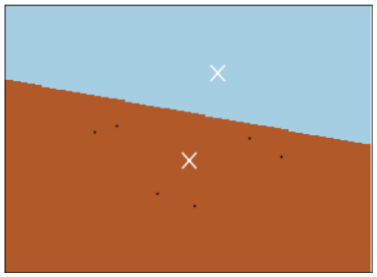

Fig. 15. PSO Clustering

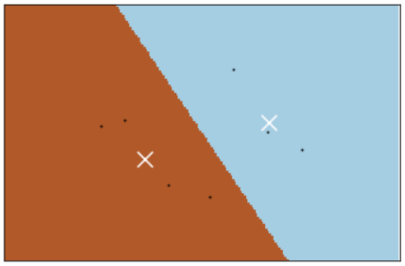

Fig. 16. Hybrid K- Mean Clustering

D. *Output for Wine Quantity Dataset*

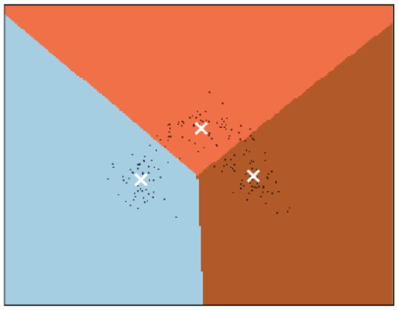

Fig. 17. K-means Clustering

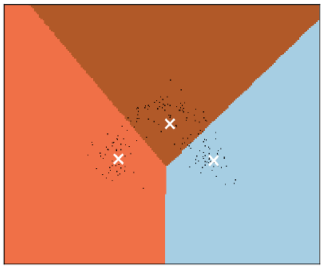

Fig. 18. PSO Clustering

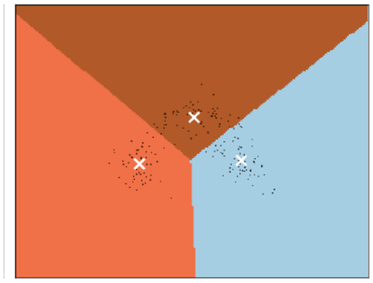

Fig. 19. Hybrid K- Mean Clustering

E. *Output for MNIST dataset*

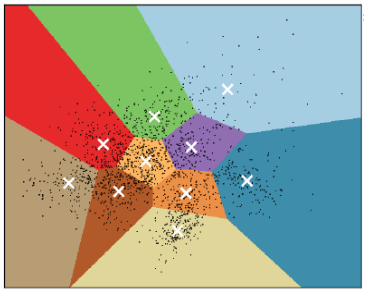

Fig. 20. K-means Clustering

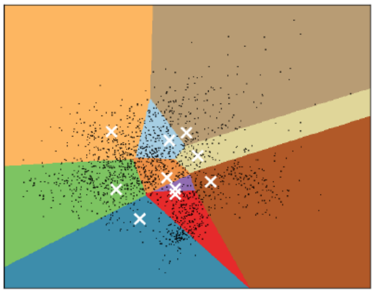

Fig. 21. PSO Clustering

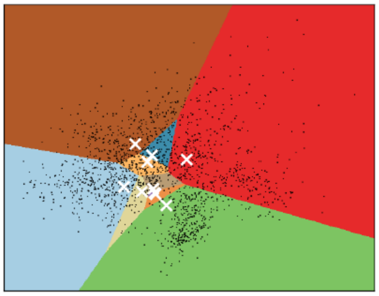

Fig. 22. Hybrid K- Mean Clustering

6. Conclusion

As a method, K methods are a quick and easy way to test data formation. However, it has its flaws, it has potential for improvement. This paper illustrated strategies to improve K-means performance through the use of first refined centers and particle efficiency. The key to unsupervised learning functionality is to understand how and when to use it. No single process will always lead to better data collection.

In the future, determining the number of collections by force using the Silhouette score may be included. Hybrid K-means algorithm courses can be expanded to cover as much detail as data integration and image integration.

References

[1] Liu, C., Wang, C., Hu, J., and Ye, Z., "Improved K-means algorithm based on hybrid rice optimization algorithm", 2017 9th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), Bucharest, Romania, 2017.

[2] Wang. J., Zhou. Y., "Particle Swarm Optimization with Generalized Local Search Operator for Global Optimization", Advanced Intelligent Computing Theories and Applications. With Aspects of Artificial Intelligence, LNCS, Vol. 4682, pp. 851-860, Springer Verlag Berlin Heidelberg 2017.

[3] Sun, J., Feng, B., Xu, W.B., "Particle swarm optimization with particles having quantum behavior", IEEE Journal Proceedings of Congress on Evolutionary Computation, 2018.

[4] Krishna, K. and M.N. Murty, Genetic K-means algorithm. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 2018.

[5] A. L. N. Fred and A. K. Jain, "Combining multiple clusterings using evidence accumulation," IEEE Transaction Pattern Intellectual, vol. 27, no. 6, pp. 835-850, 2017.

[6] Lin, Y., et al., K-means optimization clustering algorithm based on particle swarm optimization and multiclass merging, in Advances in Computer Science and Information Engineering. 2019, Springer.

[7] Neha Soni, Amit Ganatra. "Comparitive Study of Several Clustering Algorithms", International Journal of Advanced Computer Research, vol. 2, no. 6, December 2018.

[8] Yogita Rani and Harish Rohil, "A Study of Hierarchical Clustering Algorithm", International Journal of Information and Computation Technology, vol. 3, no. 11, pp. 1225-123, 2015.

[9] K. A. V. L. Prasanna and Vasantha Kumar, "Performance Evaluation of multiview-point based similarity measures for data clustering", Journal of Global Research in Computer Science.

[10] K. Sathiyakumari, and V. Preamsudha, "A Survey on Various Approaches in Document Clustering", Int. J. Comp. Tech. Appl., vol. 2 (5), 1534-1539.

[11] Anil K. Jain and Richard K. Dubes, "Algorithm for clustering dataset," 2019) prentice hall.

[12] Aurelien Geron – F, "Hands-On Machine Learning with Scikit-Learn and Tensor Flow: Concepts, Tools, and Techniques to Build Intelligent Systems," 2017.

[13] Mehryar mohair, Afshin Rostamizadeh and Ameet Talealkar, "Foundations of maching learning," 2nd edition.

[14] Peter Bruce and Andrew Vruce, "Practical Statistics for Data Scientists," 2017.