Enhance Clustering Algorithm Using Optimization

Prepared by

Dr. Anand Khandare
Associate Professor
Thakur College of Engineering and Technology
Mumbai, India

Roshakumar R. Maurya
Department of Computer Engineering
Thakur College of Engineering and Technology
Mumbai, India

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.**

I. INTRODUCTION

In unsupervised learning, the training methods do not use any forms of labels during the algorithms. This can shorten the time needed to train a classifier, and allows researchers to spot structures in the data. One of the methods in unsupervised learning is K-means method, which classifies the data into k different clusters. Each cluster is assumed to be Gaussian and spherical, with each data point belonging to the cluster whose center it is closest to.

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.

The refined clusters are then used in the K-means algorithm to classify the data. The refined initial clusters are designed to generate more predictable clustering's. Particle Swarm Optimization based clustering algorithm has been used for image and data vector clustering. This paper will be comparing the hybrid K-means algorithm against the standard PSO and standard K-means (scikit package) algorithm, over Wine, Digits dataset and also trying to use different type dataset.

II. BACKGROUND AND MOTIVATION

A. Background

Clustering is one of the challenging mining techniques in the knowledge data discovery process. Managing huge amount of data is a difficult task since the goal is to find a suitable partition in an unsupervised way (i.e. without any prior knowledge) trying to maximize the intra-cluster similarity and minimize inter-cluster similarity which in turn maintains high cluster cohesiveness. Clustering groups data instances into subsets in such a manner that similar instances are grouped together, while different instances belong to different groups.

The instances are thereby organized into an efficient representation that characterizes the population being sampled. Thus the output of cluster analysis is the number of groups or clusters that form the structure of partitions, of the data set. In short clustering is the technique to process the data into meaningful group for statistical analysis. The exploitation of Data Mining and Knowledge discovery has penetrated to a variety of Machine Learning Systems.

B. Motivation

As the amount of digital documents over the years as the Internet grows has been increasing dramatically, managing information search, and retrieval, etc., have become practically important problems. Developing methods to organize large amounts of unstructured text documents into a smaller number of meaningful clusters would be very helpful as clustering such as indexing, filtering, automated metadata generation, population of hierarchical catalogues of web resources and, in general, any application requiring document organization.

Also there are large number of people who are interested in reading specific news so there is necessity to cluster the news articles from the number of available articles, since the large number of articles are added each data and many articles corresponds to same news but are added from different sources. By clustering the articles, we could reduce our search domain for recommendations as most of the users are interested in the news corresponding to a few number of clusters.

This could improve the result of time efficiency to a greater extent and would also help in identification of same news from different sources. The main motivation is to compare different types of unsupervised algorithm to study their behaviour, advantage, and disadvantage and study how you

choose unsupervised learning algorithm based on the dataset type.
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.

III METHOD DESCRIPTION

A. Standard K-means

The refined initialization method uses a set of J subsections of the data. Each of these subsections is meant to be a random pick of a small percentage of the original data. From each subsection a set of k cluster centers is found, with any empty clusters assigned to the point with the most distortion and then re-clustering the whole subsection. Once all subsections have non-empty k cluster centers, the set of $J * k$ points are clustered using a random initialization. The result of this clustering is used as the initial centers for the K-means clustering on the whole dataset. The first was average class purity, a measure based on the labels of the data. The second was the distortion, or sum of the L2 distance squared of the data, of the clusters where L2/Euclidean distance is given as:

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

Flow Chart:

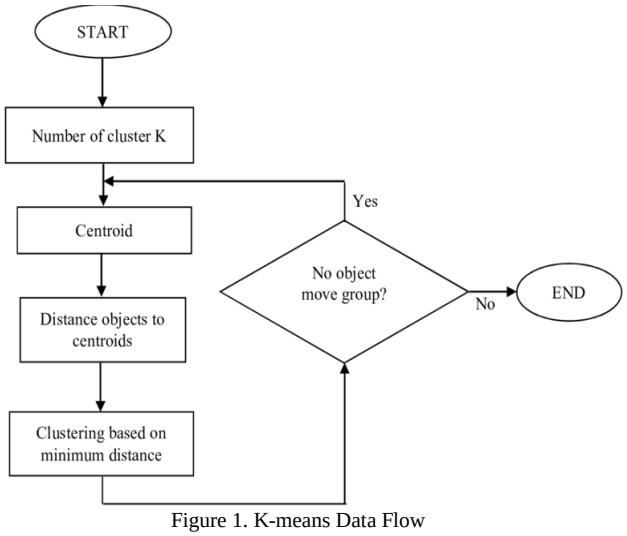


Figure 1. K-means Data Flow

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 have 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

PSO was inspired by the social behaviour of flocking of birds and originally developed by Eberhart and Kennedy in 1995 It is a population based stochastic optimization where the algorithm maintains a swarm of particles with each particle representing a solution to the optimization problem. PSO aims at finding the particle position that gives the best evaluation for a given fitness function. The following section describes the working of Particle Swarm Optimization and goes over PSO clustering and K-means PSO clustering algorithms. For this purpose the following symbols are defined:

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 components of clustering is the measure of similarity which is used for grouping data into predetermined number of clusters. Two prominent methods which are used to computer the similarity are the Euclidean distance, used for data vector clustering, and the cosine correlation measure, used for document clustering. Euclidean distance is used as similarity measure. Data vectors within a cluster are at a small 'Euclidean' distance from one another, and are associated with one centroid vector of that cluster. Distance of a vector to the centroid is determined using equation 1:

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

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:

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

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

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:

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

Flow Chart:

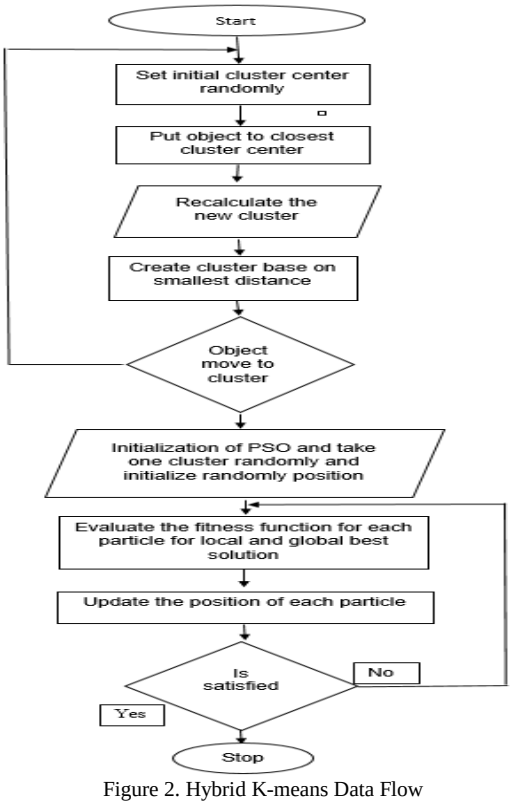


Figure 2. Hybrid K-means Data Flow

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.

IV EXPERIMENTAL RESULTS

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

In this section the results of the silhouette scores from each of the databases will be discussed. In each test, the random K-means method used was from the Scikit-learn package for Python. This was also the basis for the random initialization within the refined method. Each clustering was limited to 50 iterations. The data subsets in the refined method were each 10% of the initial dataset. The Heart Dieses Dataset was teste by varying the 300 rows within the refined dataset and this dataset hybrid k-means gives 10% more accuracy.

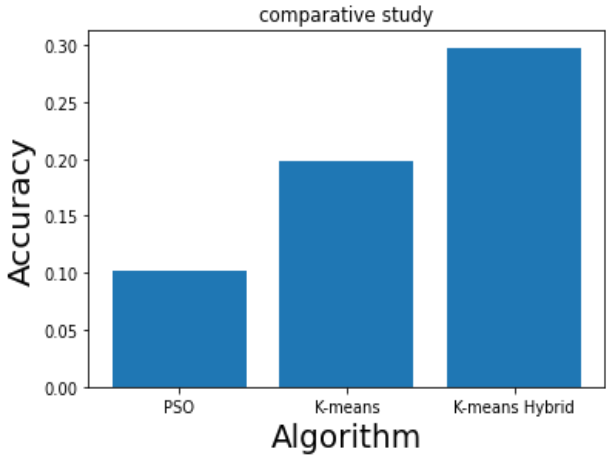


Figure 3.Heart Dieses Dataset accuracy

The number of clusters is also varied. The results show the 2 methods are comparable at their peaks. This dataset is the least complex, in that it has a limited number of both features and classes. It is this lack of complexity which can account for the similar peaks. Each method as able to identify 3 distinct classes, which matches the true characteristics. However, when the number of clusters did not match 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.

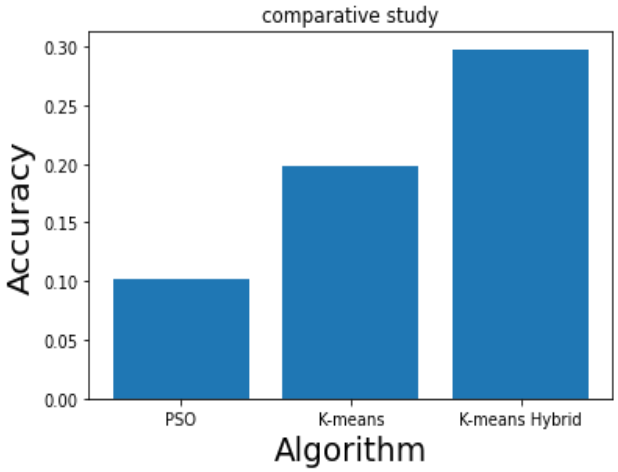


Figure 4.Breast Cancer Dataset accuracy

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

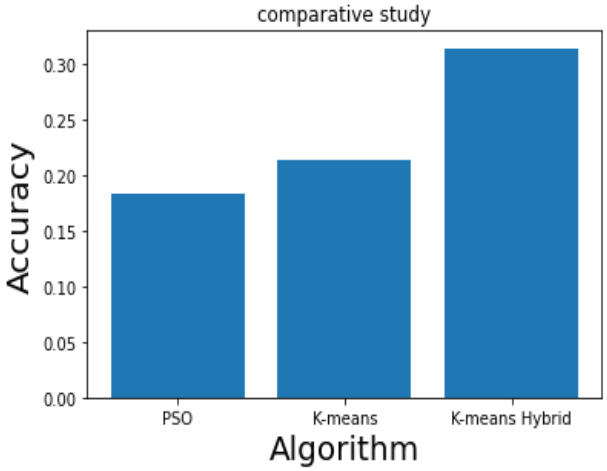


Figure 5. Diabetes Dataset accuracy

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.

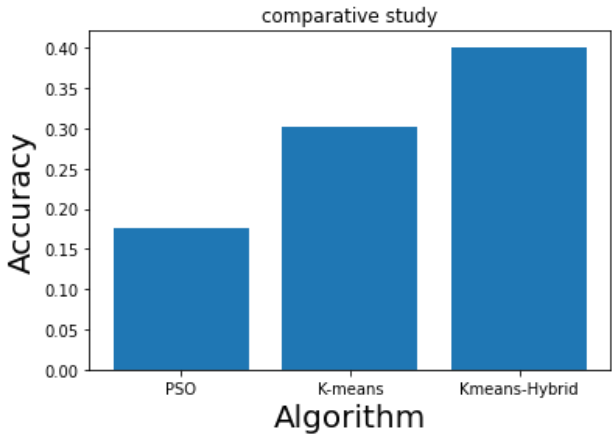


Figure 6. Wine quality dataset Accuracy

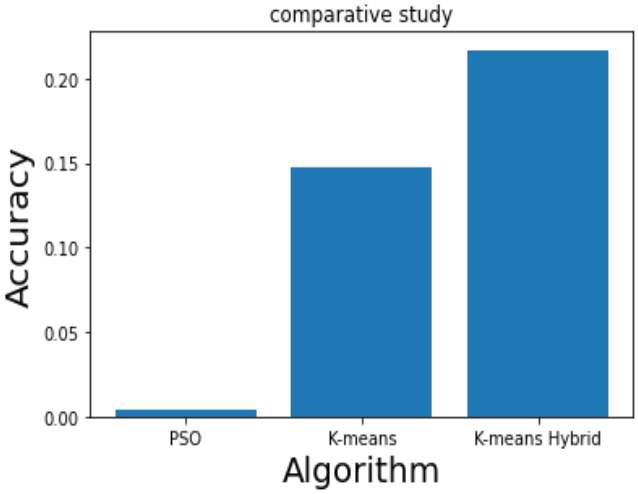


Figure 7. MNIST Dataset accuracy

This set showcases why the refined initialization method may be preferred to the random method. While there is more computation at the start of the training, it may yield better performance. This increased performance will show if the dataset is sufficiently complex, while simpler datasets may not see any improvement between the two methods.

All five dataset that show the refined method will perform as well or better than the standard k-means and PSO method given a sufficient value. The mathematical method of scoring uses the notion of distance. As features are eliminated, dimensions in the feature space are eliminated which causes smaller distances between points. However this will affect the random and refined methods equally, allowing comparison between them.
This greatly reduced feature space causes a simplified structure, leading to results. The gap between the measures of performance is the greatest from 8 to 20 features. This region shows the refined method can maintain better cluster definition as complexity increases. There is a saturation point where the methods perform similarly.

TABLE I: Algorithm Comparison

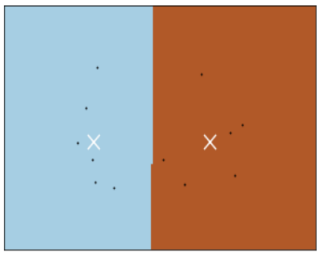
Dataset	Algorithm	Error rate	Silhouette score	elapsed time
Heart Dieses Dataset(300 rows)	PSO	1.4706	0.1022	0.0140
	K-means	1.1882	0.1976	0.0305
	Hybrid K- means	1.1751	0.2976	0.3027
Breast Cancer Dataset(600 rows)	PSO	1.4815	0.3062	0.0153
	K-means	1.0601	0.3375	0.0511
	Hybrid K- means	1.0585	0.4375	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

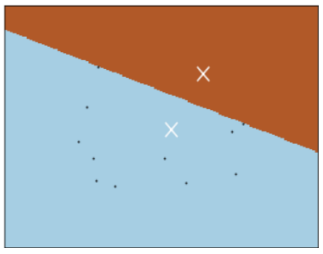
Table I lists the performance of the three algorithms on Wine and Digits dataset averaged over 10 simulations. One thing to be noted here is that although the Quantization error is comparable for the algorithms for a given dataset, it is not comparable between different datasets. This is because the quantization error depends on the number of clusters, the pre-processing of data, number of samples among other factors which vary greatly across datasets.
For the wine dataset it is clear that PSO performs worse than K-means when compared over the quantization error as well as the silhouette score. However a significant improvement can be seen in the Hybrid K-means algorithm. When one particle in PSO algorithm is seeded with the results from the K-means algorithm, the resulting algorithm performs much better than the original PSO and marginally better than the standard random K-means.

IV OUTPUTS

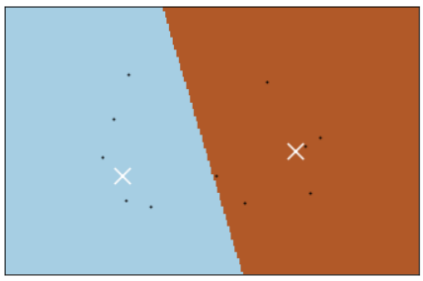
Output for Heart Dieses Dataset



K-means Clustering

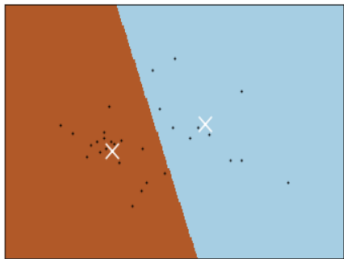


PSO Clustering

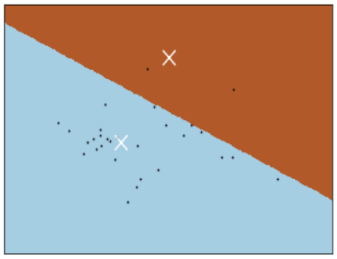


Hybrid K- Mean Clustering

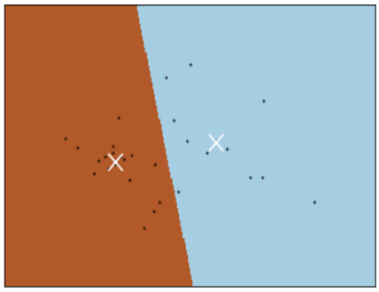
Output for Breast Cancer Diagnostic Dataset



K-means Clustering

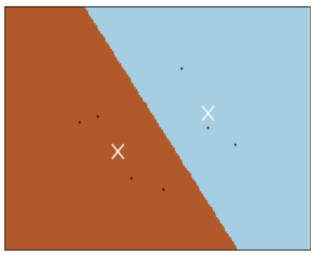


PSO Clustering

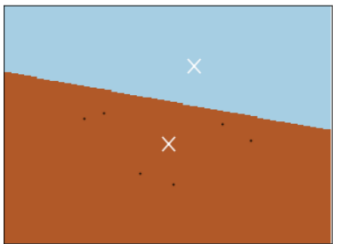


Hybrid K- Mean Clustering

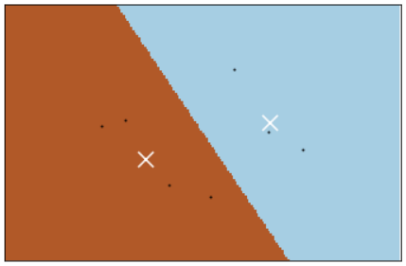
Output for Diabetes Dataset



K-means Clustering

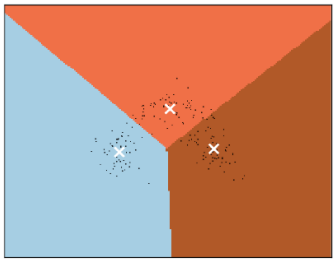


PSO Clustering

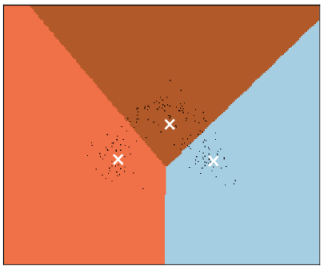


Hybrid K- Mean Clustering

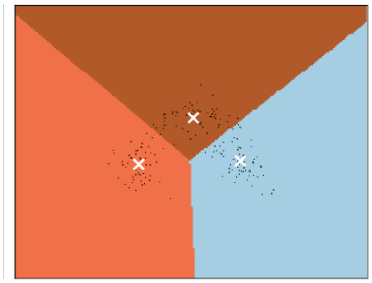
Output for Wine Quantity Dataset



K-means Clustering

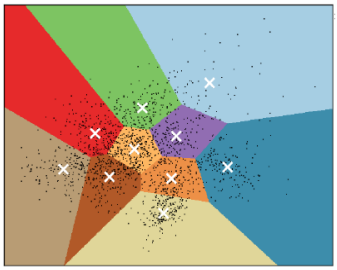


PSO Clustering

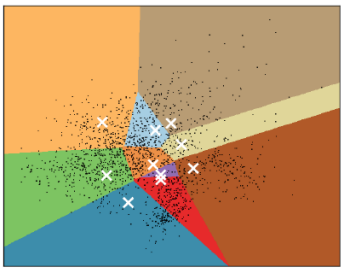


Hybrid K- Mean Clustering

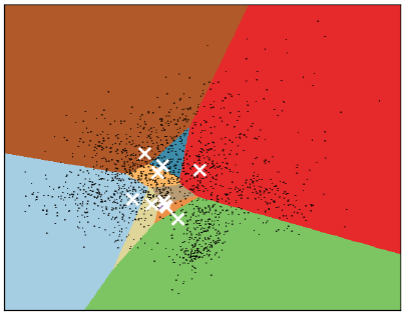
Output for MNIST dataset



K-means Clustering



PSO Clustering



Hybrid K- Mean Clustering

V. CONCLUSTION

As a technique, the K-means is a fast and simple method for examining data structures. However it does have faults, with opportunities to improve. This paper has demonstrated techniques to improve the performance of K-means through refined initial centers and particle swarm optimization. Both improvements are capable of creating less distortion while clustering data. The key to unsupervised learning performance is understanding how and when each of these should be used. There is no one technique which will always lead to the best clustering in a dataset.

In future, determination of the number of clusters dynamically using Silhouette score can be incorporated. The studies of Hybrid K-means algorithm can be extended to cover more variants of datatypes such as document clustering and image clustering.

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 usingevidence accumulation, *IEEE Transaction Pattern Intellectual*, vol. 27,no. 6, pp. 835850, 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] Prof. Neha Soni, Dr.Amit Ganatra. "Comparitive Study of Several Clustering Algorithms", International Journal of Advanced Computer Research,Volume-2,Number-4,Issue-6 December 2018.

[8] Yogita Rani and Dr.Harish Rohil, "A Study of Hierarchical Clustering Algorithm", International Journal Of Information and Computation Technology.ISSN 0974-2239 Volume 3, Number 11 (2015), pp. 1225-123

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

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