RESEARCH ARTICLE OPEN ACCESS

Experimental Study: Comparison of clustering algorithms

Mohammed Dawod[1], Mays Hasan[2], Amar Daood[2, 3]
[1] *Department of Electrical Engineering, College Of Engineering, University Of Mosul, Mosul, Iraq.*
[2] *Department of Computer Engineering, College Of Engineering, University Of Mosul, Mosul, Iraq.*
[3] *Department of Electrical and Computer Engineering, Florida Institute Of Technology, Melbourne, FL 32901, USA*
Corresponding author: Mohammed Dawod

ABSTRACT
One of the most important processes in the machine learning is the clustering. The clustering is an unsupervised process that gathers all similar measurements to identify and put them in groups based on specific measurements. Clustering task is required in many applications such as, text analysis, data visualization, nature language processing, image processing, computer vision, and even gene expression analysis. This work tends to make a comparison study to analyze the performance of different clustering algorithms using different datasets. We conduct some experimental results to evaluate the effectiveness of six clustering algorithms: hard K mean, fuzzy K mean, Locality weighted of hard K mean, Locality weighted of fuzzy K mean, Hierarchical , and DBSCAN algorithms. We use synaptic and real dataset in our experiments. We synthesize three different datasets to analyze the performance: imbalanced classes dataset, an outlier dataset, and moon dataset. Additionally, we perform image segmentation and compression using these clustering algorithms. Finally, we test the performance of the algorithms by performing facial expression clustering, which is one of the most challenging problem in the computer vision.
Keywords - Hard K mean (HKM), Fuzzy K-means (FKM), clustering, machine learning, Data mining

---

Date of Submission: 04-08-2017 Date of acceptance: 14-08-2017

I. INTRODUCTION

Clustering is the process of grouping data into clusters or groups so that all objects in the same cluster high similarity in comparison to each other, but are dissimilar to objects in other clusters. Recently, cluster analysis has been developed and improved drastically. However, there still many challenges remain unsolved. One of the main most important problem is the efficiency of clustering algorithms. Additionally, the comprehension of clustering results is a challenging problem, improving the comprehension of clustering results becomes a concern to many researchers [1].

Data clustering algorithm is based on a process of identifying the natural groupings that may exist in a given dataset, such that the objects in the same cluster are more similar and the objects in different clusters are less similar, this similarity is determined by specific measurements.

There is a wide range of clustering algorithms. Clustering algorithms can be categorized into several types, such as partitioning methods, hierarchical methods, density-based methods, grid-based methods, and model- based methods. Each one of these algorithm has strength points and limitations in terms of data characteristics that can be processed and types of clusters that can be found [2]. Clustering is a valuable tool in various applications such as pattern recognition, image processing, data mining, remote sensing, statistics, etc. [3]. As we mention before, patterns within a valid cluster are more similar to each other than patterns that are belonging to a different cluster. An example of clustering process is illustrated in Figure 1. The input patterns are shown in Figure 1(a), and the desired clusters are shown in Figure 1(b).

This paper presents a comparison study of six different clustering algorithms. We conduct several experiments using different datasets to compare the performance of these clustering algorithms. We use synaptic dataset and real dataset (Iris Fisher) in our analysis. Additionally, we use images to complete our experiments. Finally, we provide quantitative and subjective results to present the comparison.

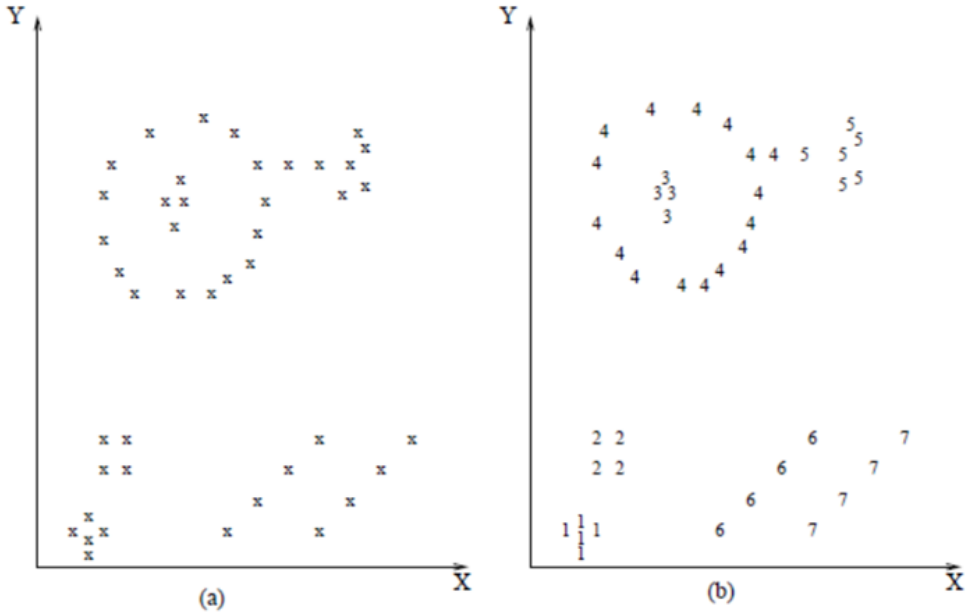


Fig. 1: data clustering (a) the original data without grouping, (b) the data with the desired groups

II. THE CLUSTERING ALGORITHMS

In this section we describe the theory behind the selected clustering algorithms.

1.1 Hard K mean

K mean is the most famous clustering algorithm. It is a very simple unsupervised learning algorithms to solve clustering problem. The procedure of this algorithm is a simple and easy way to categorize a given dataset to a certain number of clusters (K clusters which should be given to the algorithm as a priori). K mean starts by defining k centroids, one for each cluster. The next step is to take each point in the given data set and associate it to the nearest centroid. After that, the method needs to re-calculate k new centroids of the resulted clusters from the previous step. Then, a new binding is created between the points of the dataset and the new k centroids [4]. These steps are repeated in a loop. As a result of this loop we may notice that the k centroids change their location step by step until there is no more changes. In other words, centroids do not move any more. K mean aims to minimize the distance between the points of the dataset and the centroids of the clusters, its objective function is defined in the following equation:

$$j = \sum_{j=1}^{k}\sum_{i=1}^{n}\left\|x_i^{(j)} - c_j\right\|^2 \qquad (1)$$

Where k is number of clusters, n number of points, x_i are the dataset points and c_j are the centroids.

1.2 Fuzzy K mean

Fuzzy K mean (FKM) is improved version of K mean algorithm. It is developed by Dunn [5] and Bezed [6]. FKM attempts to partition the given dataset into collection of k fuzzy clusters where each data point can belong to two or more centroids. Number of centroids should be given as a prior also. FKM has similar technique to K mean except for the presence of membership for each point that gives information about the association with centroid. Instead of giving hard label to each point, the membership is calculated statistically to describe how much each point is far or close away from the centroids. The objective function of FKM is given in the following equation:

$$j = \sum_{j=1}^{k}\sum_{i=1}^{n} u_{ji}^{m}\left\|x_i^{(j)} - c_j\right\|^2 \qquad (2)$$

Where u_{ji} is the membership of x_i in the cluster c_j, and m is the fuzzy controller which determines the level of cluster fuzziness.

1.2 Locality weighted Concept

The locality weighted of K mean and fuzzy mean (LHCM, LFCM) algorithms are suggested by Huang and Zhang [7]. They used the locality information concept to improve clustering algorithms. They tried to leverage neighborhood structure information to consider the locality of some samples. Most of the clustering algorithms

treat all the data points equally and they do not give the neighborhood structure information any attention. However, some samples may contribute to the clustering results more than other samples. In locality weighted concept, an appropriate weight is given to each sample in the dataset to improve the cluster analysis. The locality weighted concept is very useful to overcome the outlier problem because it tends to assign a small weight to the outlier sample.

1.2.1 Locality weighted of hard K mean

The locality weighted of hard K mean (LHKM) is similar to hard K mean but with sensitive weighted technique. Each sample in the dataset is given an appropriate weight. The weights are calculated using the distance between the points and the centroids. The objective function of LHKM algorithm is illustrated through the following equation:

$$J = \sum_{j=1}^{k}\sum_{i=1}^{n} s_{ji} \left\| x_i^{(j)} - c_j \right\|^2 \quad (3)$$

Where s_{ji} is the weighting function to preserve neighborhood structure information of the dataset, and it is given by the following equation:

(4)

$$s_{ji} = e^{-(\|x_i - v_j\|^2)/t_j}$$

Where t_j is a scaling parameter which controls the weight matrix of the dataset points.

1.2.2 Locality weighted of Fuzzy K mean

The locality weighted of fuzzy K mean (LFKM) is improved version of fuzzy K mean algorithm. LFKM combines the membership points and the locality weighted matrix. The membership matrix gives the fuzziness to create collection of fuzzy clusters so each sample can be assigned to more than one cluster. On the other hand, the weighting function can preserve the information of the neighborhood structure and give appropriate weights to increase the significant of the close samples to the centroids. The analysis of LFKM algorithm is based on the following objective function:

$$J = \sum_{j=1}^{k}\sum_{i=1}^{n} u_{ji}^{m} s_{ji} \left\| x_i^{(j)} - c_j \right\|^2 \quad (5)$$

1.3 Hierarchical algorithm

Hierarchical clustering is introduced by Stephen in 1976 [8]. Hierarchical clustering is called by different names throughout the literatures: hierarchical, agglomerative, and tree clustering. There are two way of implementing the Hierarchical methods: top down, and bottom up. The bottom up technique starts by considering each induvial point in the dataset as a cluster. Then, the distances are measured between all clusters. These measurements are called Distance Matrix which is used to determine the closest two clusters to merge them. Then, this process is repeated until all clusters are merged together. The algorithm is described by the following steps:

1. Define each sample as a cluster
2. Compute the distance matrix among the clusters
3. Determine the closest two clusters and merge them
4. Update the distance matrix according to the new clusters
5. Repeat steps 3 and 4 until all clusters merged to create one single cluster.

Computing the similarity between two clusters in step 3 can be done using three different ways: single linkage, complete linkage, and average linkage.

1.4 DBSCAN algorithm

Density-based spatial clustering of application with noise (DBSCN) is developed by Martin Ester et al. in 1996 [9]. The main idea of this method is recognizing high density in a specific area or region and consider that area as a cluster. DBSCAN considers any area with lower density as noise because these regions should have lower density than the regions of cluster. The clusters are determined by choosing a local zone of points and minimum number of points within this zone. The zone is determined by the first parameter of the algorithm which is called epsilon. The epsilon neighborhood of data points is the set of points within a given range, the shape of that neighborhood depends on the distance between the points. The second parameter is minimum number of points inside the epsilon neighborhood. According to that analysis, the data points are divided into two groups; core points which are inside the cluster and border points which are at the border. DBSCAN starts processing the data points sequentially. It takes each point 'p' and classify it as a core point or border point according to the parameters of epsilon and minimum number of points. The point is defined as a core point when it has more than the minimum number of points within the epsilon zone. The border point has less than the minimum number of point within the epsilon and the neighborhood points should be core points.

III. EXPERIMENTAL RESULTS

In this section we describe the experimental results that we implemented to evaluate the effectiveness of the selected clustering algorithms. First, we used three artificial datasets that were generated by computer, we called these datasets imbalanced dataset, outlier dataset, and moon dataset.

The first dataset contained 150 sample points in two dimensions space, which are divided into two clusters. One of the clusters contained 25 points in a Gaussian distribution with the center (0, 0), while the other cluster contained 125 points in also a Gaussian distribution with the center (2, 2). This is an imbalanced class dataset. Figure 3 shows the output clusters of the of HKM, FKM, LHKM, LFKM Hierarchical, and DBSCAN algorithms.

The second data contained 100 sample points with two dimensions. We divided this dataset into two groups. The first group contained 50 points from a Gaussian distribution with the center (0, 0), and the second group contained 49 points from a Gaussian distribution with the center (3, 0). Additionally, we added an outlier at (200, 0) to the dataset. This dataset is an outlier dataset. Figure 4 shows the clustering results of the selected algorithms on the outlier dataset, we plotted only 99 points without the outlier for the visualization purposes.

The third dataset is shown in figure. It is non centric dataset, which consists of two interfered crescent. This kind of dataset is very challenging because it is non centric. Figure 5 illustrates the clustering results of the clustering algorithms on the dataset.

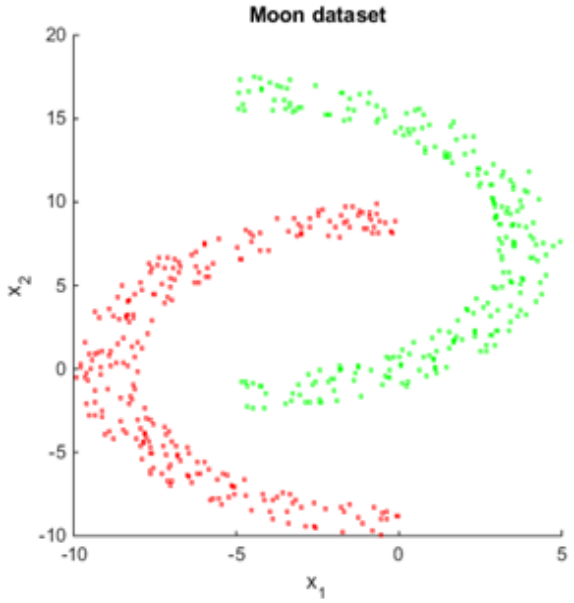


Fig.2: Moon dataset

In the second part of this work, we used image datasets to investigate thoroughly the performance analysis. The clustering process plays important role in vast range of applications such as image processing, and computer vision.

We performed image segmentation, image compression, and facial expression. Figure 6 shows the image datasets that have been used in our analysis. First, we performed image segmentation of noisy image, figure 6 (a) and (b) show the original and the noisy image. Figure 7 show the results of this experiment. Second, we performed image compression on color natural image, we used flower image as shown in figure 6 (c). We clustered the flower image to 16 clusters. The results of the compression are shown in figure 8. Finally, we perform facial expression clustering. The image dataset is shown in figure 6 (d), 70 images, where each ten images represent one of seven facial expression. The facial expressions are anger, disgust, fear, happiness, neutral, sadness, and surprise. We converted the images into row data to perform the clustering process. The results of this experiment are shown in table 4.

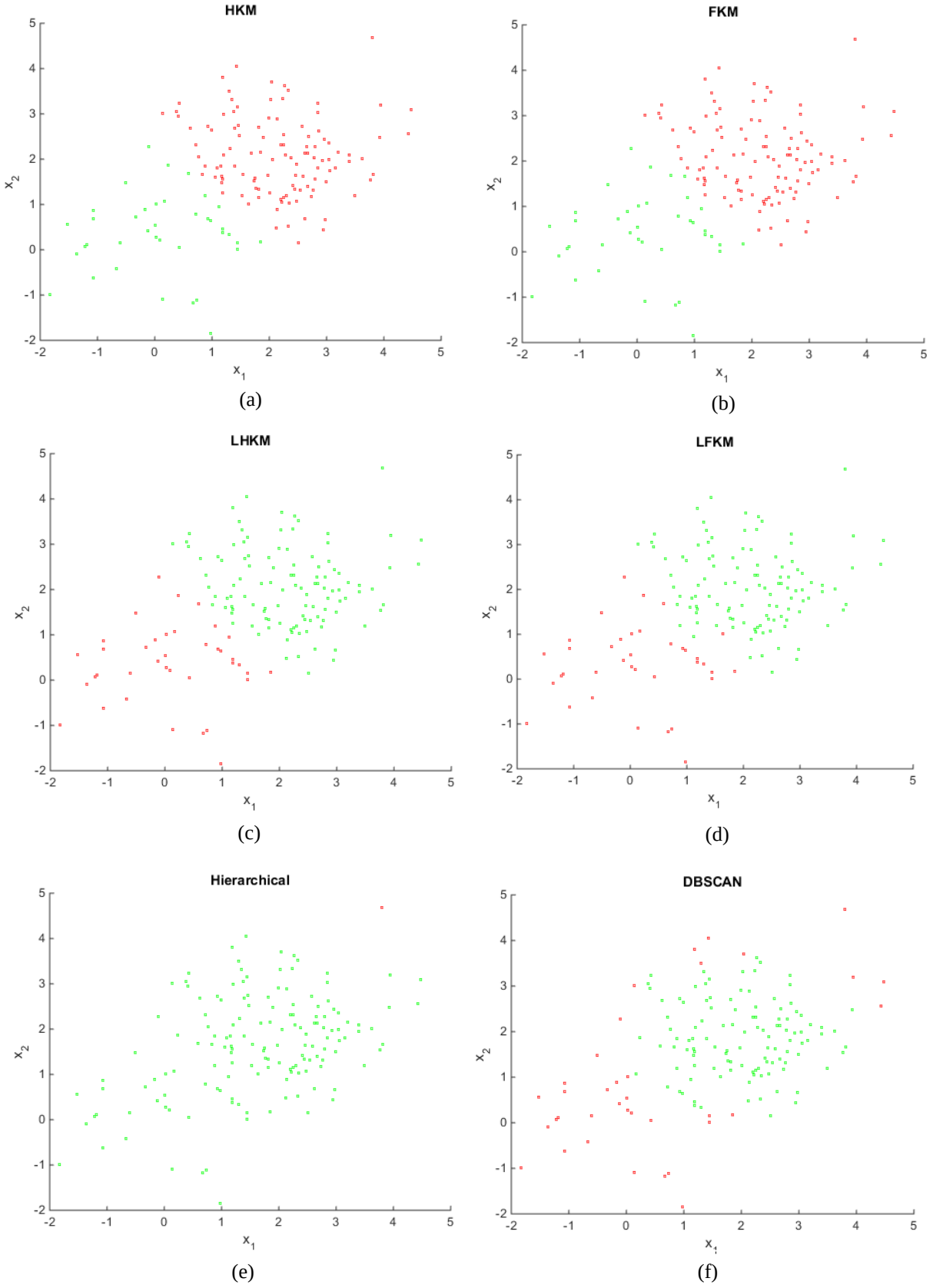


Fig.3: data clustering of the imbalanced dataset, (a) HKM, (b) FKM, (c) LHKM, (d) LFKM, (e) Hierarchical, and (f) DBSCAN

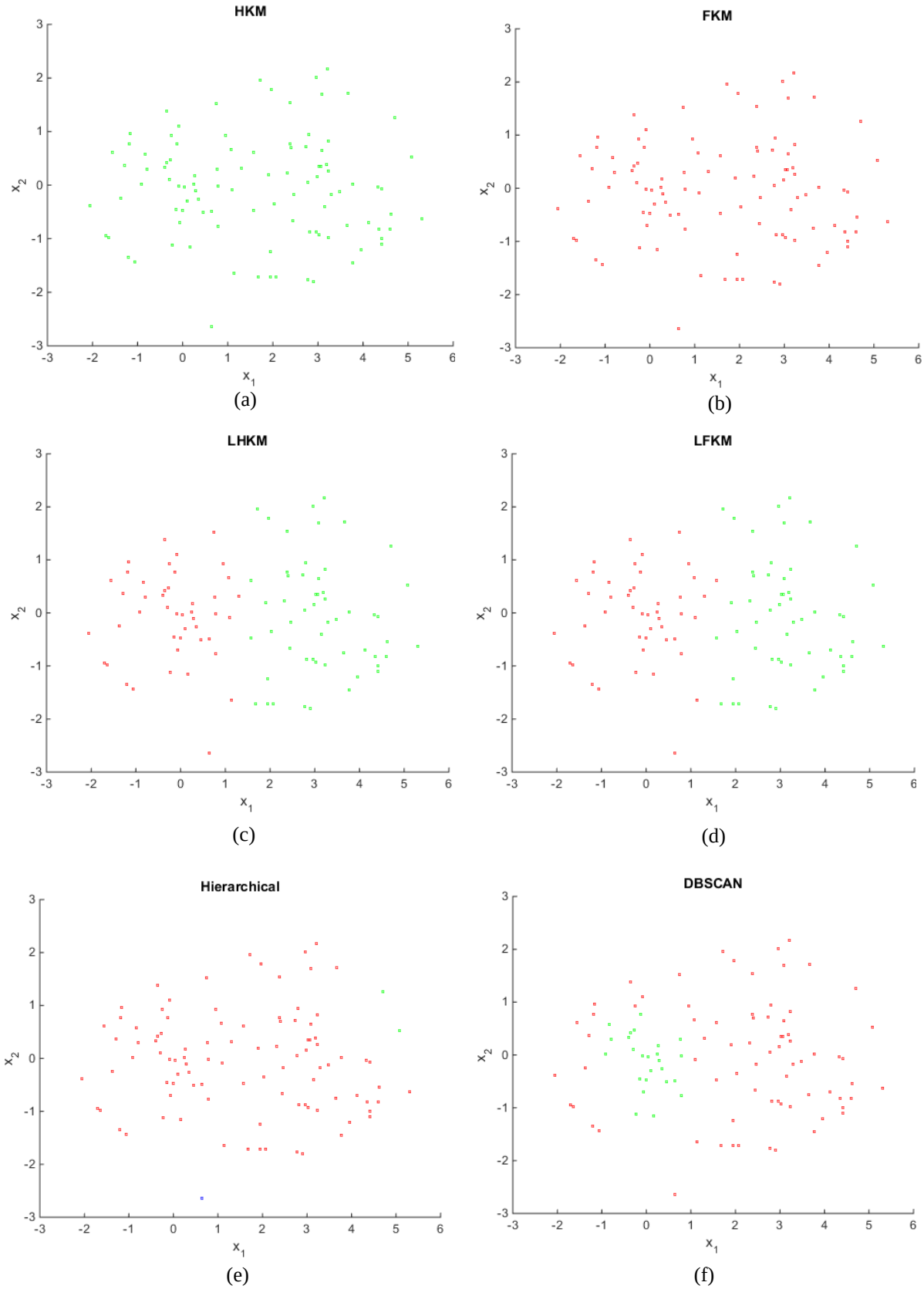


Fig.4: data clustering of the outlier dataset, (a) HKM, (b) FKM, (c) LHKM, (d) LFKM, (e) Hierarchical, and (f) DBSCAN

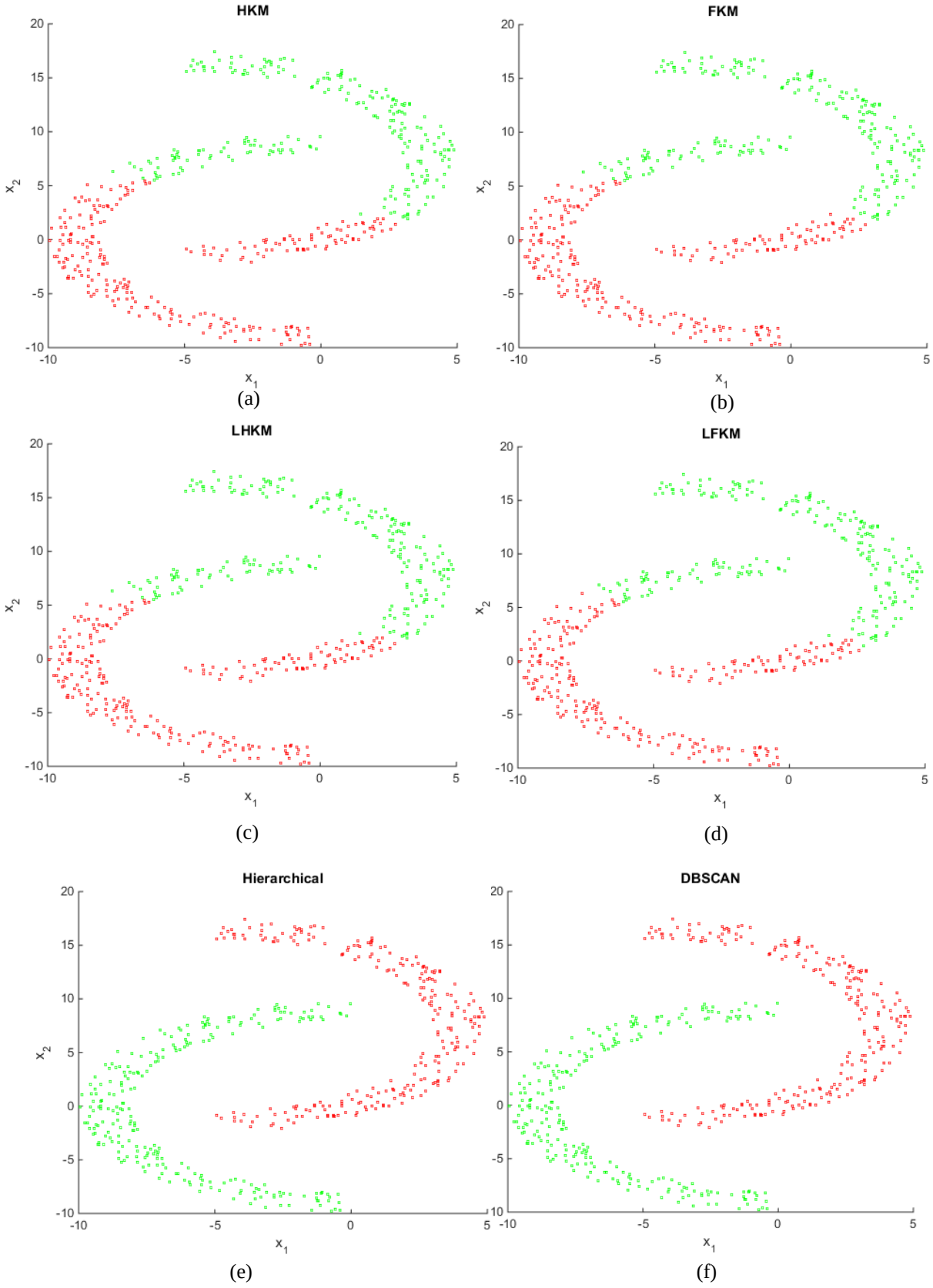
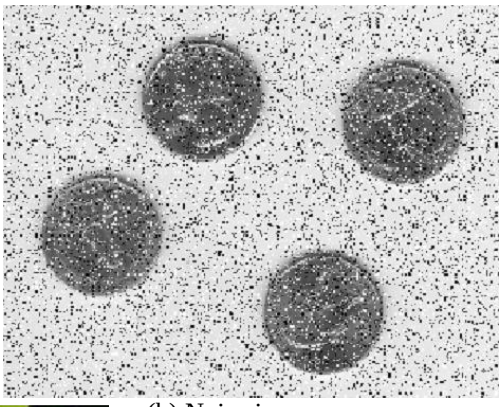
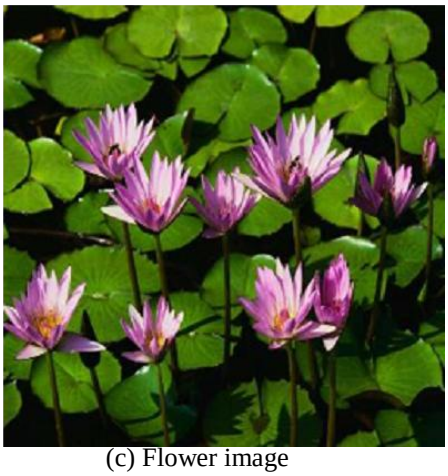


Fig.5: data clustering of the moon dataset, (a) HKM, (b) FKM, (c) LHKM, (d) LFKM, (e) Hierarchical, and (f) DBSCAN

(a) Original image



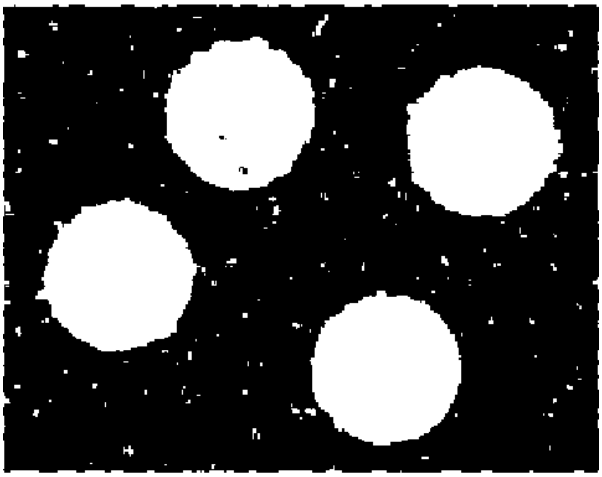
(b) Noisy image



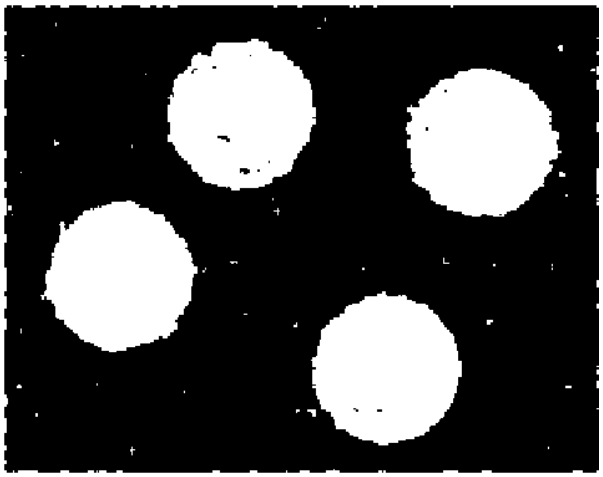
(c) Flower image

(d) Facial expression image

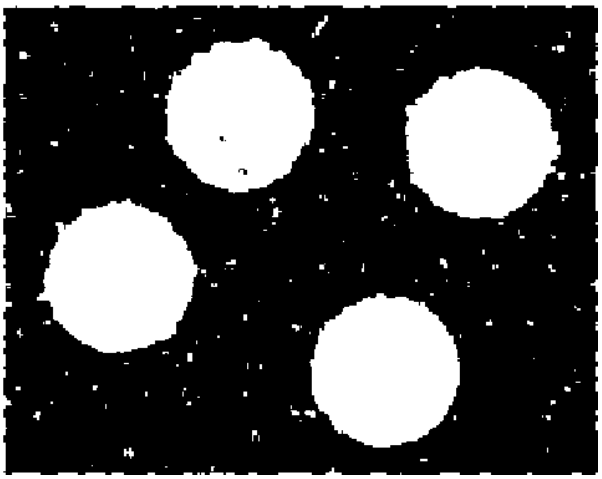
Fig.6: Image datasets, (a), (b) image dataset for image segmentation, (c) image dataset for image compression, (d) image dataset for facial expression



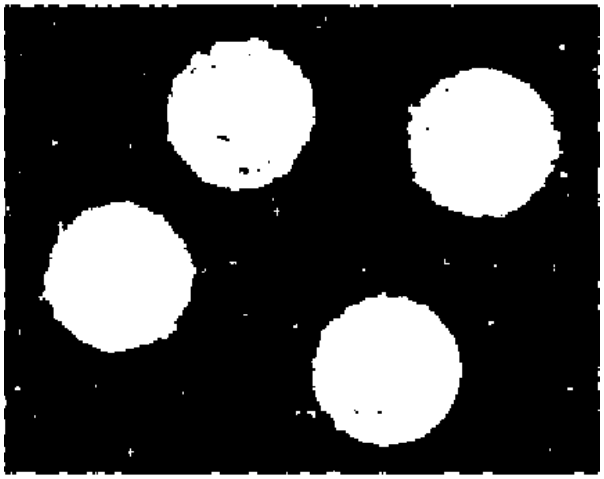
(a) HKM segmentation



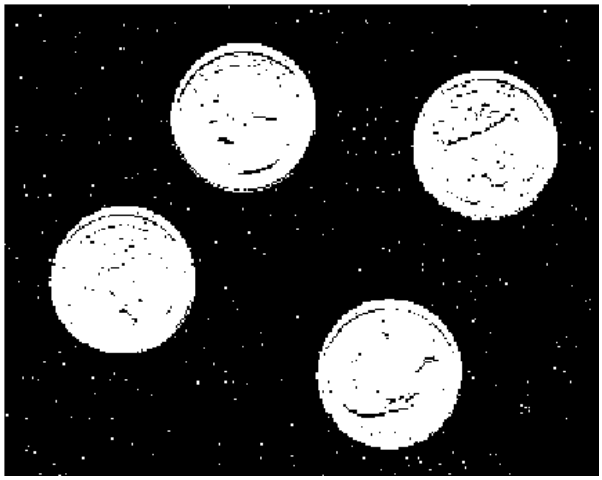
(b) HKM segmentation



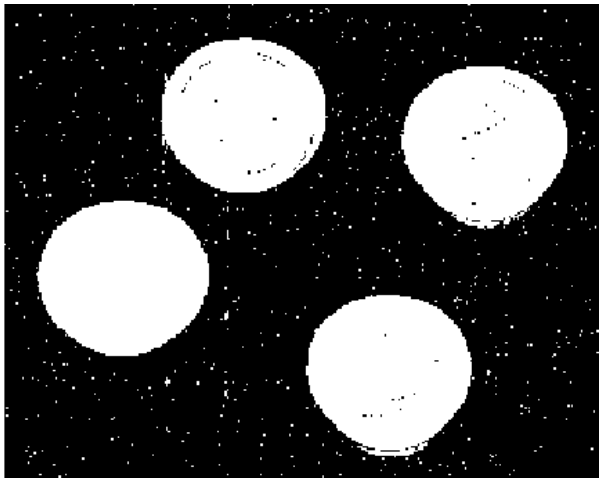
(c) LHKM segmentation



(d) LFKM segmentation

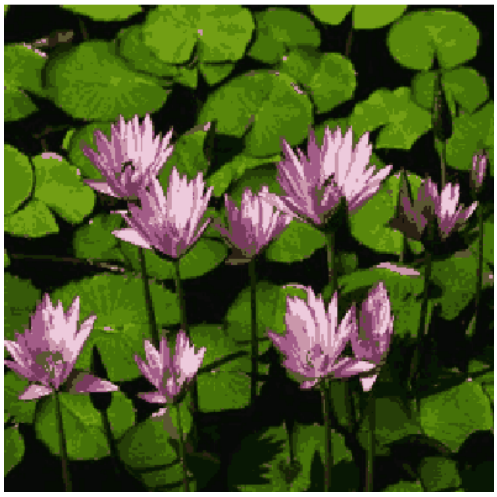


(e) Hierarchical segmentation

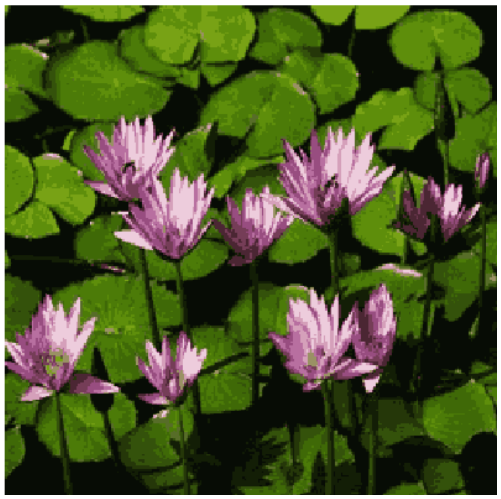


(e) DBSCAN segmentation

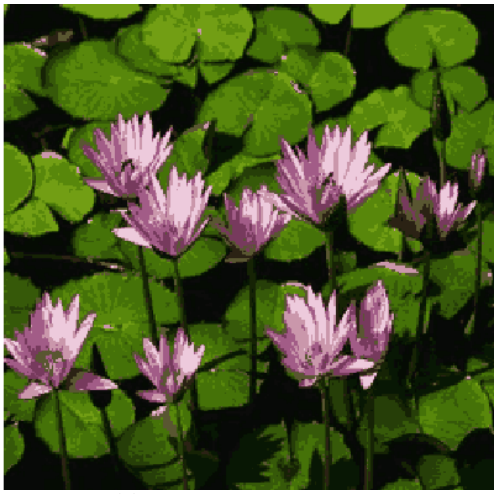
Fig.7: Image Segmentation results of the clustering algorithms.



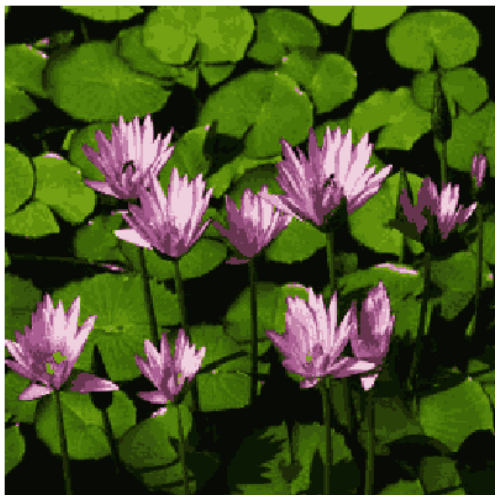
(a) HKM compression



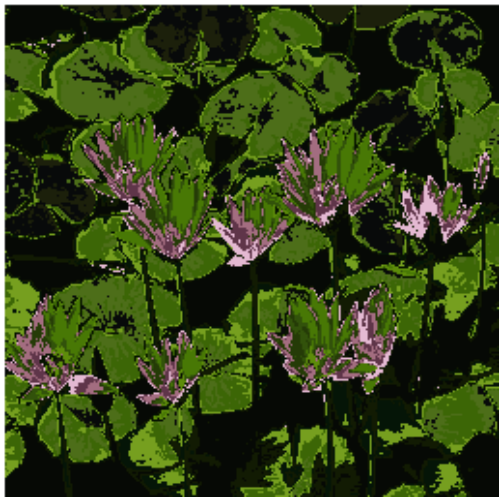
(b) HKM compression



(c) LHKM compression



(d) LFKM segmentation



(c) Hierarchical compression

(c) DBSCAN compression

Fig.8: Image Compression results of the clustering algorithms.

To compare the performance of the clustering algorithms, we used subjective and quantitative assessment. Plotting the results as shown in figures 3, 4, 5, 7, and 8 showed the subjective results. The quantitative results are presented by measurements to assess the relative performance of the clustering algorithms. There are two common way to measure the relative performance of the clustering process. The first one is called external cluster validation, this method is used when information about the true class memberships is available, but incase that we do not have this external knowledge, the second method is used which called internal cluster validation. Since, we do have the ground truth labels of our datasets, we use the external validation. We computed the precision, the recall, and the F measure to compare the performance of the clustering algorithms. These results are shown in tables 1, 2, 3, and 4.

Table 1 Imbalanced classes dataset

Method	Precision	Recall	F Measure
HKM	0.7859	0.8960	0.8212
FKM	0.7937	0.9000	0.8292
LHKM	0.8008	0.9012	0.8312
LFKM	0.8099	0.9192	0.8397
Hierarchical	0.4161	0.4960	0.4526
DBSCAN	0.7737	0.8890	0.8199

Table 2 Outlier dataset

Method	Precision	Recall	F Measure
HKM	0.4121	0.5000	0.3356
FKM	0.4121	0.5000	0.3356
LHKM	0.9537	0.9500	0.9494
LFKM	0.9623	0.9600	0.9596
Hierarchical	0.4088	0.5006	0.3487
DBSCAN	0.6111	0.6880	0.5107

Table 3 Moon dataset

Method	Precision	Recall	F Measure
HKM	0.7106	0.7100	0.7098
FKM	0.7125	0.7120	0.7118
LHKM	0.7129	0.7198	0.7121
LFKM	0.7186	0.7280	0.7178
Hierarchical	1	1	1
DBSCAN	1	1	1

Table 4 Facial expression

Method	Precision	Recall	F Measure
HKM	0.3132	0.3398	0.3154
FKM	0.4011	0.4187	0.4109
LHKM	0.4301	0.4491	0.4354
LFKM	0.4502	0.4562	0.4459
Hierarchical	0.1713	0.1677	0.1401
DBSCAN	0.2321	0.2901	0.2001

IV. CONCLUSION

In this paper, we investigated the performance of six different clustering algorithms. We used different dataset to complete our analysis. We synthesized challenging datasets to evaluate the effectiveness of the selected clustering algorithms. Additionally, we tested the performance using image datasets. We performed image segmentation, compression and facial expression clustering.

The findings show that hierarchical and DBSCAN failed to cluster the imbalanced classes and the outlier datasets. On the other hand, they performed almost perfectly to cluster the moon dataset. All the clustering algorithms have comparable results on the image segmentation. While hierarchical and DBSCAN corrupted the compressed image by performing the clustering process. Finally, the results of the facial expression were not good for all the clustering algorithms.

In the future work, we believe that, we should add more algorithms and more data to make a deeper analysis. Additionally, we should use a feature extraction to extract the facial features to improve the results of facial expression clustering.

REFERENCES

[1] TIAN Jinlan, ZHU Lin, ZHANG Suqin , and LIU Lu , Improvement and Parallelism of K-Means Clustering Algorithm, *TSINGHUA SCIENCE AND TECHNOLOGY, 10(2)*, 2005, 277-281.

[2] T. Hitendra Sarma , P. Viswanath , and B. Eswara Reddy , A Fast Approximate Kernel k-means Clustering Method For Large Data sets, *IEEE Conf. on Recent Advances in Intelligent Computational Systems (RAICS)*, 2011.

[3] A.K. JAIN, M.N. MURTY, and P.J. FLYNN, Data Clustering: A Review, *ACM Computing Surveys, Vol. 31, No. 3*, September 1999.

[4] Likas, Aristidis, Nikos Vlassis, and Jakob J. Verbeek. The global k-means clustering algorithm, *Pattern recognition 36(2)* (2003): 451-461.

[5] J. C. Dunn, A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters, *Journal of Cybernetics*, 1973, 32-57.
[6] J. C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, *Plenum Press*, New York, 1981.
[7] Huang, Pengfei, and Daoqiang Zhang, Locality sensitive C-means clustering algorithms, *Neurocomputing 73(16),* 2010, 2935-2943.
[8] Johnson, Stephen C, Hierarchical clustering schemes, *Psychometrika, 32(3),* 1967, 241-254.
[9] Ester, Martin, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu., A density-based algorithm for discovering clusters in large spatial databases with noise, *In Kdd, vol. 96(34),* 1996, 226-231.

Mohammed Dawod. “Experimental Study: Comparison of clustering algorithms.” International Journal of Engineering Research and Applications (IJERA), vol. 7, no. 8, 2017, pp. 23–34.