Unsupervised learning on Color Images

Sindhuja Vakkalagadda[1], Prasanthi Dhavala[2]

[1]*Computer Science and Systems Engineering, Andhra University, AP, India*
[2] *Computer Science and Systems Engineering, Andhra University, AP, India*
[1]*vakkalagaddasindhuja@gmail.com,* [2]*dhavalaprasanthi3@gmail.com*

Abstract

Image clustering is an evolving field with lots of applications in computer vision. The clustering result accuracy depends on the clustering algorithm, but the selection of clustering algorithm for the given image data set is a challenging problem. There are the different state of the art clustering algorithms available, but these algorithms require tuning of some parameters. The conventional clustering algorithm is K-Means. In K-Means algorithm, have to specify the number of clusters and initialization of centroids in prior, which are not known in most of the cases. The Agglomerative clustering is another popular technique, which requires the number of clusters to be specified prior and Agglomerative clustering depends on the connectivity matrix, which requires more memory to process. Another most popular clustering technique which is known for nonparametric clustering is Mean-Shift. Though Mean-Shift is a nonparametric algorithm, it does require the bandwidth parameter 'h' to be tuned. In this project, the agglomerative clustering algorithm is optimized by converting data points into the grid and graph. Mean-Shift clustering algorithm manipulated instead of static bandwidth considered plug-in rule Silverman value. All these clustering algorithms implemented using python machine learning package Scikit-learn and results are compared. The observations show that Mean-shift with Silverman bandwidth is superior to other algorithms. If properly initialize the k value k-means also working good.

.

Keywords: *Clustering, K-Means, DBSCAN, Agglomerative clustering, Mean-Shift*

1. Introduction

Unsupervised learning or clustering on images divides the image into clusters that specify a pattern between clusters. This is called unsupervised learning since there is no supervisor or teacher to guide for the division of data. Supervised learning on the other hand contains trained set of points which acts like a supervisor or teacher. Based on these trained sets, new data is tested and classified. Image clustering now a days has enormous applications in computer vision like market segmentation, medical imaging, social network analysis. The well known standard clustering algorithms are K-Means [13][10], DBSCAN [11], Agglomerative clustering [8] and Mean-Shift [3]. The conventional K-Means clustering algorithm is fast and efficient. But it depends on the number of clusters (K) to be specified. DBSCAN (Density Based Spatial Clustering on Applications with Noise) is another clustering algorithm based on the densities of data points and is robust to outliers, since all the outliers are considered as noise. The disadvantage with DBSCAN clustering algorithm is the dependency on two parameters eps and minPts. The first parameter eps is the maximum distance between two points for neighborhood check and the second parameter minPts is the minimum number of points to form a cluster.

Another bottom up hierarchical clustering is agglomerative clustering. It helps in ordering of objects. The main disadvantage with agglomerative clustering is to specify the number of Clusters like K-Means and also it

depends on connectivity matrix which consumes more memory. Mean-Shift algorithm is another clustering technique which is a data driven approach and robust to initializations. The main dependency of Mean-Shift clustering algorithm is the bandwidth selection. If the bandwidth value is higher, then result is under clustered and if the bandwidth is lesser, then the result is over clustered. Optimal bandwidth value is required for efficient clustering.

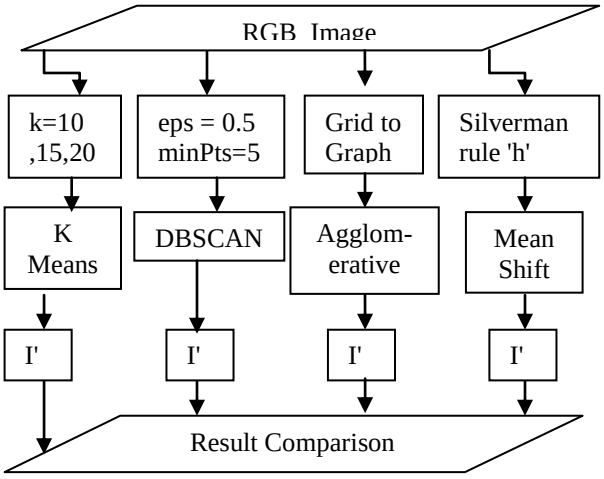

Fig 1 : Data flow diagram

For this project, the input is an RGB image which contains 3 bands: Red, Green and Blue with the values ranging from 0 to 255. Different standard clustering algorithms are applied on the image. The agglomerative clustering algorithm is optimized by converting the data points into grid for connectivity. Mean-Shift clustering algorithm is manipulated with the plug-in rule Silverman value. All these clustering algorithms implemented using python machine learning package Scikit-learn and results are compared. I' is the clustered image obtained.

2. Existing clustering Algorithms

K-Means [10] clustering algorithm is one of the popular conventional clustering algorithms. The K-Means clustering algorithm takes the input K (The number of clusters to form) and refines K number of clusters based on a similarity measure. The similarity measure mostly used is Euclidean distance. The initial step is assignment of K centroids which are chosen or generated randomly. Then each data point is assigned to any of K centroids based on the Euclidean distance. In the next step the cluster centroids are updated to the mean of all points in that cluster. This is an iterative process until no change of data points between clusters or till a predefined maximum number of iterations. K-Means clustering algorithm takes less time to execute, but the biggest disadvantage to specify the number of clusters K prior to its execution.

Density Based Spatial Clustering of Applications with Noise (DBSCAN) [11] is another graph based clustering algorithm which mainly depends on two parameters called eps and minPts. The first parameter 'eps' is the minimum distance between two points for checking their neighborhood. The second parameter minPts specifies the minimum number of points to form a cluster. Find the neighbor points within eps for all the points and identify the core points. The core points are the points with more than minPts neighbors. Then find the connected components in the neighbor graph for all core points. Non-core points are assigned to near cluster if it is in eps distance. If a data point doesn't belong to any of the clusters, then it is considered as noise. DBSCAN is robust to outliers because of noise. Also it is independent of number of clusters specification like K-Means. But the main disadvantage is that it depends on eps and minPts parameters.

Agglomerative hierarchical clustering [8] is a bottom up approach and works by merging the data one by one based on the nearest distance measure like Euclidean distance. This merging procedure is stopped when a single cluster is formed. The synthesis starts at level 0 (L(0), sequence number m=0) by taking all the data points as disjoint clusters. Then find the least distance pair (let us say r, s) over all pairs and increment the sequence number m. Merge the pair (r,s) and follow the same for the next clustering. This level is set to L(m) and update the distance matrix. This procedure is iterated till all the data points form

a single cluster. The main disadvantage with agglomerative clustering is to update the distance matrix and also it depends on the number of clusters to form. The distance matrix size becomes N^2 in the first step if the data size is N.

Mean shift [4] considers feature space of data points as a probability density function. Mean shift considers the input data points are sampled from the formed probability density function. The dense regions (or clusters) present in the obtained feature space correspond to the local maxima (or mode) of the probability density function. We can also identify clusters associated with the given mode using Mean Shift. For each data point, Mean shift associates it with the nearby peak of the dataset's probability density function. For each data point, Mean shift defines a window around it and computes the mean of the data point . Then it shifts the center of the window to the mean and repeats the algorithm till it converges. After each iteration, we can consider that the window shifts to a more denser region of the dataset. At the high level, we can specify Mean Shift as follows : 1) Fix a window around each data point. 2) Compute the mean of data within the window. 3) Shift the window to the mean and repeat till convergence.

3. Parameter optimization

In hierarchical agglomerative clustering to find connected components of core points, it maintains a connectivity matrix. The connectivity matrix is updated while execution. But initially the size of connectivity matrix becomes N^2 if the actual data size is N. To store such a big matrix is difficult and required more space. So got memory error while executing the algorithm. To overcome this problem, the data points are fitted to a grid and converted to graph. Because there is no necessity of storing the whole connectivity matrix at a time, since we required only connected components with minimum distance.

Modified Agglomerative clustering :

1) Each point as seperate cluster
2) Fit the points to a grid and to graph, Prepare a distance matrix
3) Take nearer points and combine into a cluster
4) Take each cluster as a single point
5) Update the distance matrix
6) Repeat the process till a single cluster formed

As mentioned earlier, Mean-Shift clustering algorithm required bandwidth parameter to be specified even though a data driven approach. There is no global bandwidth for Mean-Shift since the feature space of data points changes with various images. There are some standard rules called Silverman Rule of thumb and Scott rule. In this paper some static bandwidths, Scott rule bandwidth and Silverman bandwidths are computed and applied.

Scott's Rule of Thumb is a plug - in rule for evaluating bandwidth is defined as

$$h = 1.059 * A * n^{-1/5}$$

where,
A = min(std(sample), IQR/1.349)
IQR = Inter Quartile Range
std(sample) = Standard Deviation of Sample points
n = Number of sample points

Silverman Rule of Thumb is another standard plug-in rule for calculating bandwidth and it is defined as

$$h \approx 1.06 * \sigma * n^{-1/5}$$

where
σ = Standard deviation of sample points
n = Number of sample points

The h value obtained from the above plug-in rules is given as bandwidth 'h' for Mean-Shift algorithm.

Modified Mean-Shift :

1) Calculate Silverman bandwidth 'h'
2) Fix a window around each point within 'h' distance
3) Compute the mean
4) Shift window to the mean
5) Repeat till convergence

Comparison is done between Mean-Shift clustering with static, Scott and Silverman bandwidths along with other optimized clustering techniques.

4. Results

K-Means clustering algorithm required the parameter K to be tuned. For comparison the K values are taken as 10, 15 and 20 clusters. For 15 clusters, we got 0.69 as Silhouette score. Also for 15 and 20 clusters, the Silhouette scores are 0.75, 0.70 respectively. The main disadvantage for K-Means clustering is to define the value K. If K value is known prior, K-Means clustering will give better results. For abbreviating K-Means clustering simply, I have used the notations of K10 for K-Means with 10 clusters, K15 for K-Means clustering with 15 clusters and K20 for K-Means with 20 clusters. There are some methods to find the value of K like Elbow method and Silhouette valued graph method. But those are applicable when we interpret or assume the range of values for K.

DBSCAN on the other hand required eps and minpts parameters to be tuned. In this project default values of 0.5 and 5 for eps and minpts respectively are taken from sklearn library. Results scored 0.65 Silhouette value. But those default values are suited for the image that is considered. If the image changes, then the performance is also changing.

Mean-Shift clustering technique is dependent on bandwidth parameter. There is no chance of having a global value for bandwidth since it changes with the data set. The Elbow method and Silhouette valued graph methods can also be used in Mean-Shift to determine the value of bandwidth. But before that we should know the range of values where the bandwidth value resides. For results comparison, fixed bandwidth values of 5, 10 are taken. Also as mentioned above, bandwidth values are calculated and considered from Scott and Silverman rules. Among all those bandwidth values, Mean-Shift clustering with Silverman bandwidth is giving better performance with 0.80 Silhouette value.

Fig 2: Actual Image

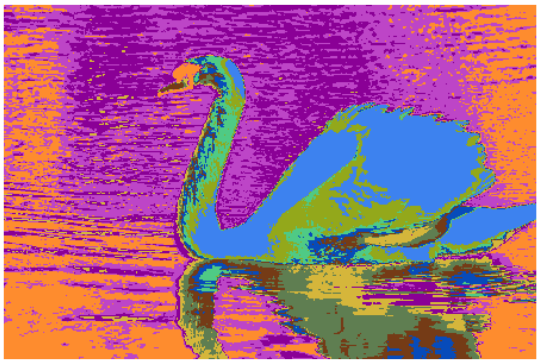

Fig 3: K-Means clustering with 10 clusters

Fig 4 : K-Means clustering with 15 clusters

Fig 5: K-Means clustering with 20 clusters

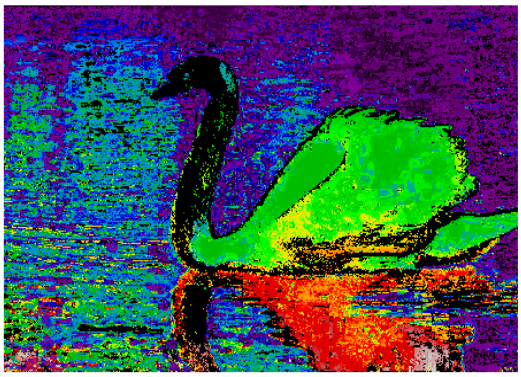

Fig 6 : DBSCAN clustering

Fig 7 : Agglomerative clustering with 10 clusters

Fig 8 : Agglomerative clustering with 5 clusters

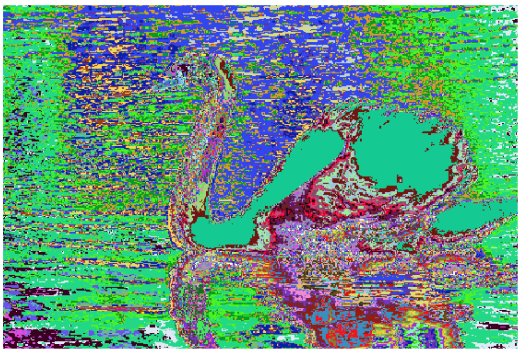

Fig 9 : Mean-Shift with static bandwidth h = 5

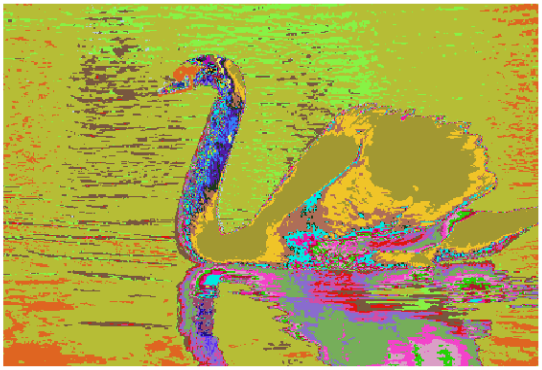

Fig 10 : Mean-Shift with static bandwidth h = 10

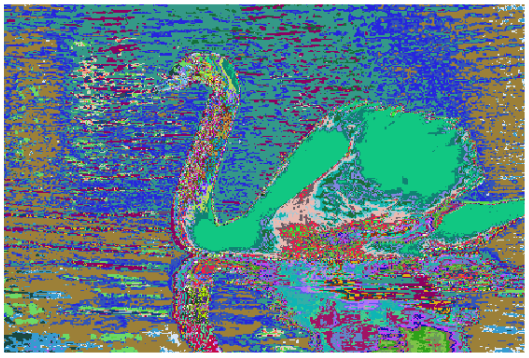

Fig 11 : Mean-Shift with Scott Rule bandwidth

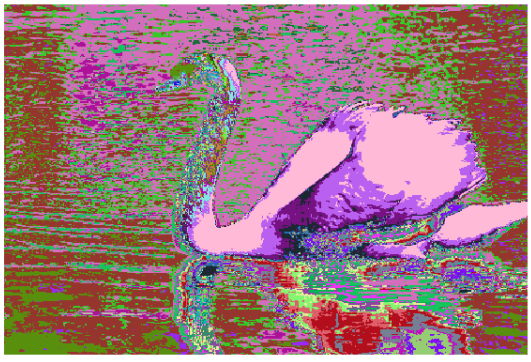

Fig 12: Mean-Shift with Silverman bandwidth

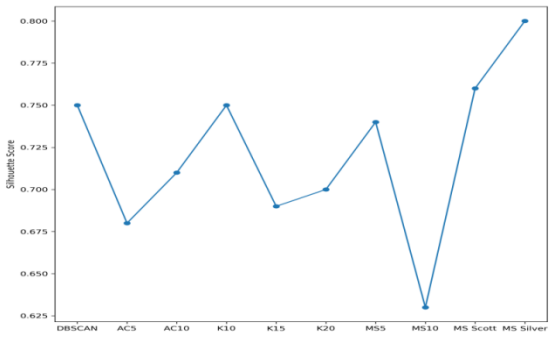

Fig 13 : Silhouette values graph

5. Conclusion and future work

Every clustering algorithm has its own advantages and disadvantages. Also most of the clustering algorithms have dependency on parameters. K-Means and Agglomerative clustering rely on the number of clusters to be specified in prior. The agglomerative clustering algorithm optimized by converting data points into the grid. Mean-Shift clustering algorithm requires less parameter tuning. The standard rules for bandwidth evaluation are helpful in increasing efficiency of Mean-Shift clustering algorithm.

One of such plug-in rule is Silverman rule of thumb that is performing better with Mean-Shift clustering algorithm compared to other clustering techniques. Still selection of k and initialization of centroids of K-means and DBSCAN parameters eps for neighborhood check and minPts for minimum number of points in a cluster must considered properly. These parameters evaluation and selection of relevant clustering algorithm for the given image data set can be solved as future work.

Bibliography

[1] B. Zhou, J. Pei and W. Luk, " A brief survey on anonymization techniques for privacy preserving publishing of social network data", ACM SIGKDD Explorations Newsletter, vol. 10, no. 2, pp. 12-22, 2008.

[2] Biplab Banerjee, Surender Varma G, Krishna Mohan Buddhiraju , "Satellite Image Segmentation: A Novel Adaptive Mean-Shift Clustering Based Approach ",Ieee Igrss,2012.

[3] D. Comaniciu and P. Meer, "Mean-shift: A robust approach toward feature space analysis ". IEEE Trans. Pattern Anal. Machine Intell., 24:603–619, 2002.

[4] Johannes Jordan, Elli Angelopoulou , "Meanshift Clustering For Interactive Multispectral Image Analysis".

[5] ZHAO Yunji1, PEI Hailong2, LIU Baoluo3 , "Mean-shift algorithm based on kernel bandwidth adaptive adjust ".

[6] Sangeeta Yadav , Mantosh , "Improved Color-Based K-mean Algorithm for Clustering of Satellite Image ".

[7] Rokach, Lior, and Oded Maimon. "Clustering methods." Data mining and knowledge discovery handbook. Springer US, 2005. 321-352.

[8] Zhang, et al. "Graph degree linkage: Agglomerative clustering on a directed graph." 12th European Conference on Computer Vision, Florence, Italy, October 7–13, 2012.

[9] Ester, Martin; Kriegel, Hans-Peter; Sander, Jörg; Xu, Xiaowei (1996). Simoudis, Evangelos; Han, Jiawei; Fayyad, Usama M., eds. A density-based algorithm for discovering clusters in large spatial databases with noise. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96).

[10] Hamerly, Greg; Drake, Jonathan (2015). "Accelerating Lloyd's algorithm for k-means clustering". Partitional clustering algorithms:

[11] "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise" Ester, M., H. P. Kriegel, J. Sander, and X. Xu, In Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, 1996

[12] "Mean shift: A robust approach toward feature space analysis." D. Comaniciu and P. Meer, IEEE Transactions on Pattern Analysis and Machine Intelligence (2002)

[13] "Web Scale K-Means clustering" D. Sculley, Proceedings of the 19th international conference on World wide web(2010)