OPEN JOURNAL SYSTEMS
ISSN:2237-2202

Available on line at Directory of Open Access Journals
Journal of Hyperspectral Remote Sensing v.10, n.2 (2020) 117-121

www.periodicos.ufpe.br/revistas/jhrs

Journal of Hyperspectral RemoteSensing
www.ufpe.br/jhrs

Study of the Clustering Algorithms for Hyper Spectral Remote Sensing Images

Addanki C. Reddy[*], Addanki Saraschandrika[**], Addanki V.Reddy[***]

[*]Bachelor of Technology in IT, 4thyear, SRM IST, Kattankulathur, Chennai, 603203. Email: chandrahasreddy03@gmail.com(Corresponding author)
[**]Bachelor of Technology in Civil, 4th year, IITRAM, Ahmedabad, 380026
[***]Department of Physics, JawaharBharathi Degree College, Kavali, 524201

Received 23 June 2020; accepted 21 October 2020

Abstract

The data taken from the hyperspectral images are discrete and hard to classify because they are arranged in the contiguous spectral bands. We can easily detect and classify the data from the spectral images if the number of attributes in the images is very little. But it is very difficult to segregate the data from the images if the numbers of classes are more. To make the segregation easy we implement the procedure that utilizes a clustering algorithm. This paper comprises of two sections, firstly to perform unsupervised learning using different types of clustering algorithms and secondly, to compare the efficiency of the resultant clustering of these different methods to prove that which clustering method is best suitable in reading the hyperspectral imaging data. For this I have used these clustering algorithms, they are DBSCAN, MiniBatch K-Means, K-Means. By comparing these techniques I surmised that the K-Means is better for using the HyperSpectral Imaging data. To perform these calculations I used the Matlab data set from the Computational Intelligence Group.
Keywords:clustering, machine learning, hyper spectral imaging

1. Introduction

For the past decade, the geological survey has been done completely by humans. It comprises of a lot of work including the surveying the whole land done by one team and collecting the data from them and separate and combine the identical classes by one team, and finally creating the raw map of the piece of land by another with all those details. It may even take more than 20 days for a single piece of land. If the land is too big for example amazon forest it may take months to complete the whole process. But with the advancement of computing and technology, the whole process will take less than a day. With the help of the satellites, we can take the Hyperspectral images and with machine learning algorithms we can segregate the imaging data in a fraction of seconds.

In this analysis, we have taken the data from the Computational Intelligence Group providing the imaging data over the Salinas Valley in California. The data was collected by the Airborne Visible / Infrared Imaging Spectrometer 224-band sensor. This image data was aimed at classifying the 16 samples in that Salinas land. It includes Broccoli green weeds, Fallow lands, Stubble, Celery, Grapes, Soil vineyard, corn and lettuce of different kinds(Lee e Kwon, 2017).

The Hyperspectral Imaging (HSI) data is to express the image in the spectrum for each spectral so that it can identify different objects like hills, lakes, food corps, roads, etc., The ability of this peculiar hyperspectral imaging is that our human eye is capable of understanding three wide-range spectral colors they are red, blue, green (Bilgin et al., 2008**)**. For a given piece of land even if the similar materials are wide apart, with the assigned spectrum we can identify them easily. This spectral imaging is a 2-D monochromatic output. It uses a continuous range of wavelengths with 400nm to 1100nm. The main applications of the HSI are used in agriculture, food processing, surveillance, chemical imaging, etc. The main advantage of using the HSI is that the whole image is obtained once and the individual need not to have any prior knowledge of the sample (Zhang et al., 2016).

Clustering is a type of unsupervised learning algorithm, in which the input data is unlabeled. Usually, it is used to obtain a meaningful output from the set of unlabeled data. The main definition of clustering is to form the set of points in a group usually cluster. For the task consists of a huge amount of data, the clustering algorithm makes sure that all the points in the data must be aligned to a certain cluster and no point should be left. From these, we can classify the data based on the number of clusters and similarly we can train the data based on these clusters so that the additional data must be formed in this cluster or formation of the new cluster.

DBSCAN is the data based clustering algorithm, in which it establishes the spatial clustering based on the density of the points that outlines the low-density points to the high-density points(Hou et al., 2016).The main advantage of the density-based spatial algorithm is that there is no need to mention the number of clusters.This is the updated version of the K-Means algorithm, in which it iterates the data and used that dataset to update the clusters of the class. This process is iterated until convergence. The clusters are been updated providing a convex combination of data. K-Means is the most popular method of the vector quantization, the main function of this algorithm is that it separates the data points by initiating the number of clusters to be assigned and arranges the data by utilizing the shortest path from the centroid(Wang et al., 2019). Mostly, it uses Gaussian shortest distance formula to calculate it. Then the whole process is iterated until the whole data points are been assigned to the particular cluster.

This comparative model for implementing and analyzing the different clustering algorithms for hyperspectral imaging. The papers that are previously presented are only based on either the hyperspectral data or the clustering algorithms. Nobody has proposed the comparative analysis of different clustering algorithms regarding the data using hyperspectral data.

This paper proposes a different algorithm known as a spectral-spatial sparse clustering algorithm for the HSI. This proposed algorithm has taken the rich spatial information in the SSC model to improve the performance of the clustering algorithm. Even though it is different from others, it still needs certain improvements so that this model can be improved automatically determining the cluster detection (Zhang et al., 2016).

This paper proposes a letter that suggests a new algorithm L2-SSC algorithm by concluding the spatial contextual information. This experiment concludes that clustering performance for the Hyperspectral Images. The advancement that can be made in this algorithm is to improve the efficiency and performance of the architecture (Zhai et al., 2017).

This paper presents the DBSCAN as the parameter, and propose to use histogram equalization algorithm. In his experiments, they got that the DBSCAN algorithm can be used as an effective image segmentation approach. Hence their algorithm focuses on the overall structure of images and tends to generate large segments (Hou et al., 2016).

This paper proposes the unsupervised LDA and can be transformed into K-Means while the subspace is fixed, this proposed un-LDA optimizes in each iteration and continuously increasing until it converges, but this is not sensitive to the regularization parameter, but with good performance (Wang et al., 2019).

This paper proposes a letter that suggests HSI classification based on fuzzy-clustering algorithms, this proposes a novel approach to include spatial information in the segmentation process by mentioning 2 and 3-D Gaussian filtering of fuzzy-membership degrees(Bilgin et al., 2008). This provides the best clustering performance for hyperspectral images.

The main objective of the paper is to determine which unsupervised algorithm is suitable for reading and manipulating the data. Hence we implemented three different clustering algorithms and compare those efficiencies in finding out which algorithm would be the best choice for depicting the hyper spectral data.

2. Methodology

To minimize the time and the labour work to reconstruct the data taken from the hyperspectral remote sensing images. We planned to create an algorithm that analyzes the data taken from the images and after certain trials; we came to know that the classification algorithm doesn’t work better for spectral imaging data. So, we planned to use the clustering algorithm because based on the data points that been distributed in the graph we can identify the exact shape of the image along with the clustering of the similar objects that provides the monochromatic output of the spectral image.

To compute the clustering algorithm for the data, we have used the dataset taken from the Computational Intelligence Group. They obtained the data AVIRIS sensor providing with 224-band. It consists of 512 lines by 217 samples as the dataset for the ground truth(Lee e Kwon, 2017). The main ideology behind this hyperspectral image is to identify the agricultural-related corps. The dataset taken from the CIG is in .mat files that we reorganized it into the python code to get the clustering output.The proposed system comprises of four modules. The overview of the model with four modules is presented below in Figure 1.

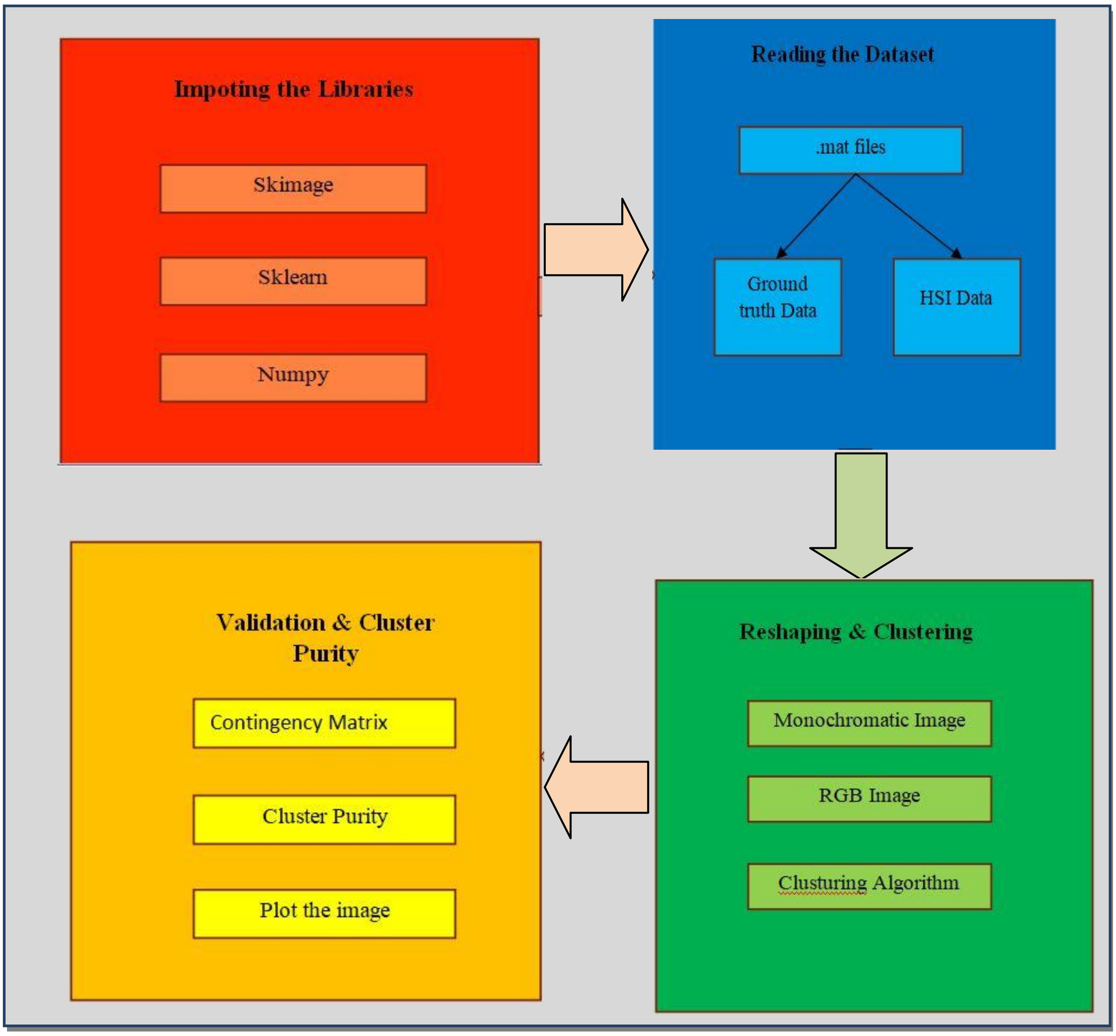

Figure 1- Overview of the proposed model.

Importing the libraries.

As we have used the clustering algorithm in the python, it is a mandatory condition to import the libraries to execute the code without any errors. The libraries that we have used are scipy, which is used to load the Matlab dataset to the python, secondly, NumPyie., Numerical Python that helps in calculating the purity of the cluster, Thirdly, sklearn, this is the major step used to import the different clustering algorithms for our code. Finally, skimage, this is used to reshape the obtained monochromatic output into RGB colored image.

Reading the dataset

The data that we have taken from the CGI is in the .mat format. So, with the normal code, we can't import the dataset, so with the help of the scipy library, we have loaded that Matlab dataset to my code. In this, we have used two Matlab files because one is the ground_truth data and the other is the corrected data. In which the ground_truth data consists of 512 lines for the 16 classes of the Hyperspectral data. This also includes the transformation of the data matrix to the array similar to the width and height of the 2-D matrix.

Reshaping and Clustering

This step includes the reshaping of the matrix of the corrected data to the tuple. Also reshaping of the matrix data to the recognised image array so that it can be suitable for the clustering algorithm. For the analysis we have used three different clustering algorithms with various functional parameters. This step also includes the conversion of the monochromatic image output to the coloured image for better visual perception that is labelling into RGB conversion.

Validation and Cluster Purity

In this step primarily we create the contingency matrix also known as confusion matrix, this can be done with help of the sklearn library. In which it creates the matrix by considering the truth value and the predicted value. Then finally, we obtain the efficiency by calculating the division of max of the confusion matrix to the sum of the contingency matrix. With this we can get the purity of the data for the Hyperspectral image.

3. Analysis

Upon verifying the hyperspectral imaging data taken from the CGI(Lee e Kwon, 2017), we plan to find the agricultural sector in that area. To find that from the data we first plan to do it using supervised learning, but upon verifying the data we came to know that supervised learning is not suitable for the HSI data(Zhang et al., 2016). Then we tried to use the clustering algorithms. Upon varying the variables we found that the clustering is best suited for the HSI data. Then we implemented a python code that produces the cluster purity of the algorithm. In order to do that first, we have to produce the Figure 2 monochromatic output of the data.

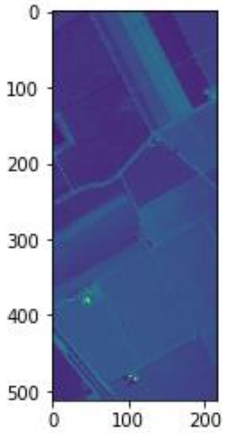

Figure 2- Monochromatic image of DBSCAN.

By seeing this image output it is very difficult to identify similar pieces of land and the proportion of land. So with the help of the skimage library, we implemented the code to produce the colored image of Monochromatic output in Figure 3(a) and Figure 3(b). Then we got the result for both the MiniBatch K-Means and the K-Mean as follows.

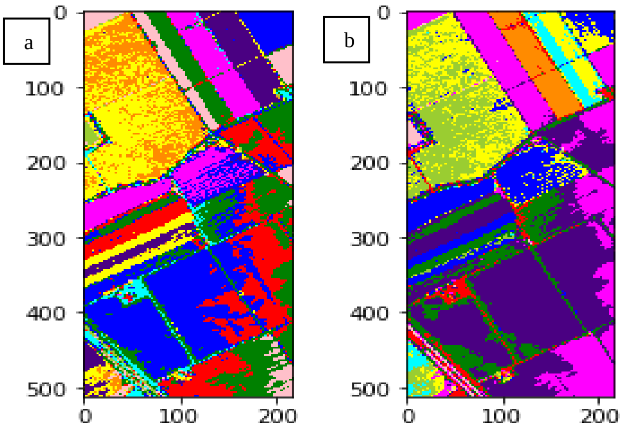

Figure 3 - (a) K-Means and (b) MiniBatch K-Means.

From these two figures, we can easily identify which parts of the image are of similar colors along with the borders of the road in between the two fields. Then we tried to compare the algorithms for the same dataset and classes. Then we got that K-Means has the highest cluster purity with 67.29% efficiency, secondly MiniBatch K-Means with 65.46% and finally DBSCAN with 20.82% efficiency. Figure 4 shows the graphical representation of the cluster purity of the three clustering algorithms. Where, cont_mat is contingency matrix.

cont_mat = contingency_matrix(y_true, y_pred)

cluster purity = np.sum(np.amax(cont_max, axis = 0)) / np.sum(cont_mat)

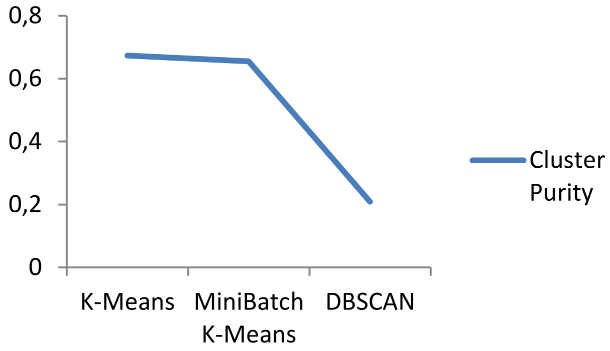

Figure 4 -Analysis of Cluster Purity.

4. Conclusion

This paper proposes the analysis of the performances of the three different clustering algorithms that are DBSCAN, MiniBatch K-Means, and K-Means, by utilizing the dataset provided by the CGI on the agricultural piece of land in Salinas valley in California. The data that we have taken are Hyperspectral remote sensing images. These data are in .mat format. With the help of these, we analyzed the performance of the image in order to find which algorithm is better for reading the hyperspectral data. By computing the data with various algorithms and by analysing the cluster purity of those algorithms, we came to know that K-Means is the best algorithm for the study of geological data by HSI. Secondly, MiniBatch K-Means and lastly DBSCAN, it shows the less efficiency in reading the data by our cluster purity algorithm. The analysis result may differ if we change the number of clusters for the algorithm. For the future study if there is any other algorithm that is best suited for the clustering especially for the hyperspectral images can be encouraged. This paper is only analyzed based on three clustering algorithms, remaining algorithms are for future study. Finally, if there is any technology that automatically implements the purity score along with the HSI will reduce the time complexity of the whole process.

References

Bilgin, G., Ertürk, S.,Yıldırım, T., 2008. Unsupervised classification of hyperspectral-image data using Fuzzy approaches that spatially exploit membership relations. IEEE Geoscience and Remote Sensing Letters 5, 673-677. doi: 10.1109/LGRS.2008.2002319

Hou, J., Gao, H., Li, X., 2016. DSets-DBSCAN: a parameter-free clustering algorithm. IEEE Transactions on Image Processing 25, 3182-3193. doi: 10.1109/TIP.2016.2559803

Lee, H., Kwon, H., 2017. Going deeper with contextual CNN for hyperspectral image classification. IEEE Transactions on Image Processing 26, 1-1. doi: 10.1109/TIP.2017.2725580

Wang, F., Wang, Q., Nie, F., Li, Z., Yu, W., Wang, R., 2019. Unsupervised Linear Discriminant Analysis for Jointly Clustering and Subspace Learning. IEEE Transactions on Knowledge and Data Engineering. doi: 10.1109/TKDE.2019.2939524

Zhai, H., Zhang, H., Zhang, L.,Li, P., Plaza, A., 2017. A new sparse subspace clustering algorithm for hyperspectral remote sensing imagery. IEEE Geoscience and Remote Sensing Letters 14, 43-47. doi: 10.1109/LGRS.2016.2625200

Zhang, H., Zhai, H., Zhang, L., Li, P., 2016. Spectral–spatial sparse subspace clustering for hyperspectral remote sensing image. IEEE Transactions on Geoscience and Remote Sensing 54, 3672-3684. doi: 10.1109/TGRS.2016.2524557