doi:10.20944/preprints202106.0634.v1

Unsupervised Classification of Hyperspectral Images using PCA and K-Means

Ayesha Malik
Department of Electrical Engineering
Institute of Space Technology
Islamabad, Pakistan
ayesha.malik17@ist.edu.pk

Mamoona Waheed
Department of Electrical Engineering
Institute of Space Technology
Islamabad, Pakistan
mamoona17@mail.ist.edu.pk

***Abstract*—The visualization of hyperspectral images in display devices, having RGB colour composition channels is quite difficult due to the high dimensionality of these images. Thus, principal component analysis has been used as a dimensionality reduction algorithm to reduce information loss, by creating uncorrelated features. To classify regions in the hyperspectral images, K-means clustering has been used to form clusters/regions. These two algorithms have been implemented on the three datasets imaged by AVIRIS and ROSIS sensors.**

Keywords—Hyperspectral image, HSI, PCA, K-means clustering, unsupervised, classification, bands, satellite, ROSIS, AVIRIS

I. Introduction

Remote Sensing is the data collection of earth's surface, by measuring the reflected signal at various wavelength bands to form a spectral signature from objects at long distances, via either of the two techniques: infrared sensing or optical remote sensing (via satellite's optical sensor) [1]. Based on the number of bands, satellite images are classified into: panchromatic, multispectral and hyperspectral images.

Attaining hyperspectral image datasets depends on imaging spectrometers that image in continuous, narrow regions of electromagnetic waves [2]. Due to which, they have high spectral resolution that describes boundaries and composition of an object and can help in image classification, which is the most active part of research. Classifying hyperspectral images may seem difficult due to their high dimensionality, missing labels, variation in data due to spatial locality and poor image quality due to noise or background interference etc.

Hyperspectral image classification methods are as such [3]: supervised, unsupervised and semi-supervised. Supervised classification is based on prior knowledge. An example of this method is the support vector machine (SVM). On the contrary, unsupervised classification is not based on any prior knowledge of data. A common method is K-means clustering and its aim is to obtain a meaningful output from the set of unlabelled data via clustering/grouping [4].

Hyperspectral imaging (HSI) is used for [5]: agricultural and water resources control (e.g. identifying crops, lakes etc. [4]), document imaging and mineralogical mapping of earth surface.

The large hyperspectral data, increases processing complexity and time. The solution is to effectively reduce data to apply various processing methods e.g. classification, that will help find the number of regions in hyperspectral image [6].

The remainder of this paper is structured as follows. Section 2 presents the literature review of work done in the field. Section 3 defines the problem statement. Methodology is explained in Section 4. Dataset used and classification results form Section 5. Finally, concluding remarks alongside future work are given in Section 6.

II. Literature Review

Prior to remote sensing, geological surveys were done manually. Now, technological advancements help capture hyperspectral images via satellites and enable algorithms to help classify the data in a matter of moments [4].

Colour images have intensity values in three bands (red, green and blue), whereas, hyperspectral images have more than 100 bands in discrete intervals from visible to infrared region [7]. HSI obtains the spectrum for each pixel in the image frame to identify target areas.

Hyperspectral images are characterized by two resolution types, namely [5]: spatial and spectral. Spatial resolution is the relationship among image pixels and is inversely proportional to patch size i.e. smaller the patch, greater the details. Whereas, spectral resolution is defined as the number of bands and range of electromagnetic spectrum measured by the sensor.

High dimensional hyperspectral data lacking labelled samples can result in Hughes phenomenon i.e. reduced accuracy in classification. This can be overcome by dimensionality reduction techniques e.g. Principal component analysis (PCA), Discrete wavelet transform and Independent component analysis (ICA) [1]. The first few principal components of PCA result in 70 percent correct classification rate [6].

Previously, only spectral information was being used to achieve hyperspectral image classification upon high dimensional data via support vector machine (SVM) , neural networks etc [3]. Whereas, some classification approaches work on reducing high-dimensional data first and then using clustering methods.

Supervised deep classification of hyperspectral images produces almost 100 percent accurate results for datasets [8]. Classification methods are based on parametric and non-parametric classifiers or supervised and unsupervised classifiers [9]. Each having its own limitations [10]. Classifying regions help study the geographic location and boundaries of a certain patch that is imaged by the satellite.

Existing work is related to either PCA based schemes or discriminant analysis-based schemes. One of the prominent work done is based on both of these schemes, that is by using PCA for dimensionality reduction and SVM for classification [11].

SVM, a supervised learning method, is effective in high dimensional spaces and is accurate for classification. It requires a smaller number of training samples and labels [1]. However, SVM is computationally expensive. Thus, unsupervised classification methods such as K-means clustering , K- nearest neighbour, Neural networks are a better solution where pixels are to be assigned to clusters without using prior knowledge [12]. K-means accounts for 78.3% accuracy for classification [13].

III. Problem Statement

During hyperspectral data analysis, the following challenges are faced:

- High computational cost
- Storage of a few hundred megabytes required
- Redundancy due to correlation of neighbouring image bands. This high dimensional redundant data needs to be catered for, to efficiently apply processing techniques [14].

The aim of this research paper was to use a machine learning algorithm to find the number of regions present in each hyperspectral image and colour label to classify different regions

IV. Methodology

To achieve the goal of this paper, two-stage process was implemented i.e. first, data dimensionality was reduced, followed by K-means clustering. PCA is a standard dimensionality reduction algorithm for hyperspectral images [12] that aims to remove the correlation among bands [1].

Clustering techniques focus on grouping pixels by iteratively updating centroid of a group, where the centroid represents the mean spectral signature of all the pixels in a cluster [8].

A. Importing libraries in Python to load and process data

- numpy: for multidimensional arrays and matrics
- scipy: to load .mat file (hyperspectral dataset)
- sklearn: to import PCA and K-means
- matplotlib: to create figures and plots
- spectral: to use functions for hyperspectral data processing
- import h5py: to import data of HDF5 binary format

B. Dimensionality Reduction via PCA

PCA is a classical, unsupervised dimensionality reduction method which does not require labels [14] and preserves most of the spectral information in a compact number of principal components [1]. It is based on the fact that neighboring bands of hyperspectral images are highly correlated and often convey almost the same information. Most of the information may only be contained in the first few bands [6].

The basic principle used in PCA is the eigenvalue decomposition of the covariance matrix [1]. The eigenvalues represent the variance in the direction of the eigenvector.

It must be remembered that PCA is a dimension reduction tool, not a classifier. Thus, a classifier is fit onto the PCA-transformed data.

To choose the number of principal components needed to describe the data, explained variance ratio graph was plotted (number of components vs. cumulative explained variance). Components which had greater than 95 percent cumulative explained variance were chosen i.e. 8 principal components in this case for all three hyperspectral datasets, as shown in Fig. 1 for Pavia. The curve depicts how much of the total, 102-dimensional variance is contained within the first N components. The same approach was used to select principal components for Cuprite and Moffett.

C. K-means clustering

K-Means being an unsupervised method, groups data into 'k' number of user defined clusters. To help decide the number of clusters, elbow method was used i.e. for each value of k (e.g. in the range 1 to 10) , the sum of squared distances from each data point to its centre(distortion) was calculated.

Elbow is the point of inflection on the curve which is the optimal value of k. According to this method, k=5 for Pavia (as shown in Fig. 1) and Cuprite data sets, k= 6 for Moffett [3].

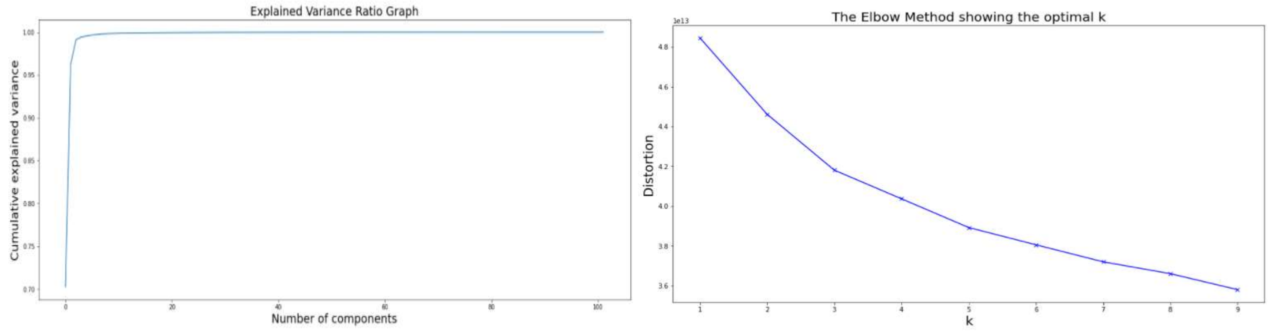


Fig. 1. Explained Variance and Elbow graph for Pavia dataset

V. Experimentation and Results

In accordance with the problem statement, three hyperspectral images were utilized [1]. The classification of different regions was done for the following datasets acquired by their respective sensors:

A. ROSIS(Reflective Optics System Imaging Spectrometer) Pavia Centre: covers Pavia, northern Italy having 102 spectral bands.

B. AVIRIS (Airborne Visible Infrared Imaging Spectrometer) datasets:

1) Cuprite: covers Cuprite in Las Vegas, U.S having 224 channels.

2) Moffett Field: covers Moffett airfiled, CA, U.S. having 224 channels.

Unsupervised classification was applied on the above stated hyperspectral images using python libraries. The datasets have the following dimensions i.e. rows, columns and bands, respectively:

- Pavia: 1096*715*102
- Cuprite: 512*614*224
- Moffett: 753*1923*224

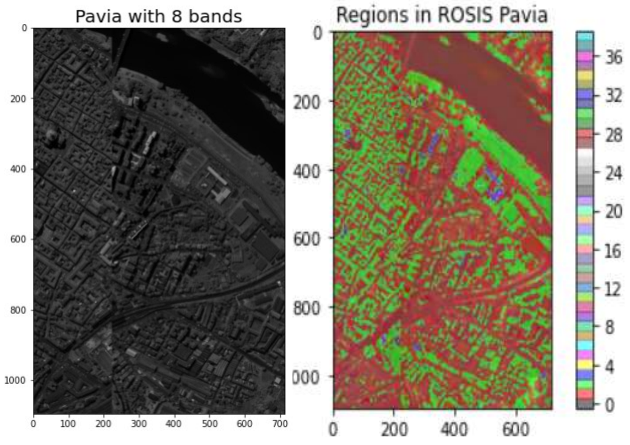
The reshape command was used to reshape the data, by multiplying the actual rows and columns to form the current rows and the original bands were retained as columns e.g. Pavia was reshaped to 783640*102.

To visualize the explained variance ratio, its cumulative sum was plotted which helped select the number of principal components to be retained as shown in Fig. 2 (a), Fig. 3 (a), Fig. 4 (a), that display the images with 8 principal components.

After reducing the dimensions of all three datasets, k-means was applied using the elbow method to assign ‘k’ optimal clusters. Fig. 2 (b), Fig. 3 (b) and Fig. 4 (b) are the colour-labelled clusters formed for Pavia, Cuprite and Moffett, respectively. The axes of Fig. 2 to Fig. 4 represent the rows and columns of the respective images in 2D.

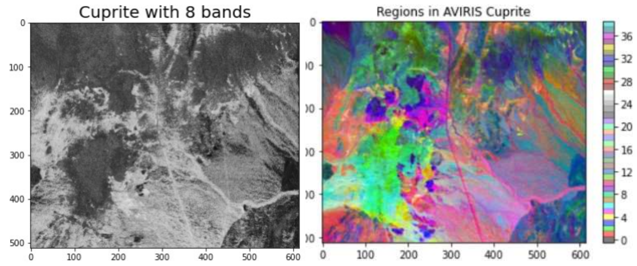
The ‘KMeans’ command used in python formed clusters with equal variance to minimize ‘within cluster sum of squares’ (WCSS). For data with high dimensionality, Euclidean distances increase. Thus, implementing a dimensionality reduction algorithm like Principal component analysis (PCA) before K-means clustering can help overcome this problem.

The number of clusters/regions formed for Pavia are 4 after 15 iterations, whereas, for Cuprite and Moffett there are 5 clusters after 15 iterations. In simple words, the algorithm segments the image by clustering pixels into classes based on the spectral similarity of pixels e.g. roads, trees, water bodies, soil etc.



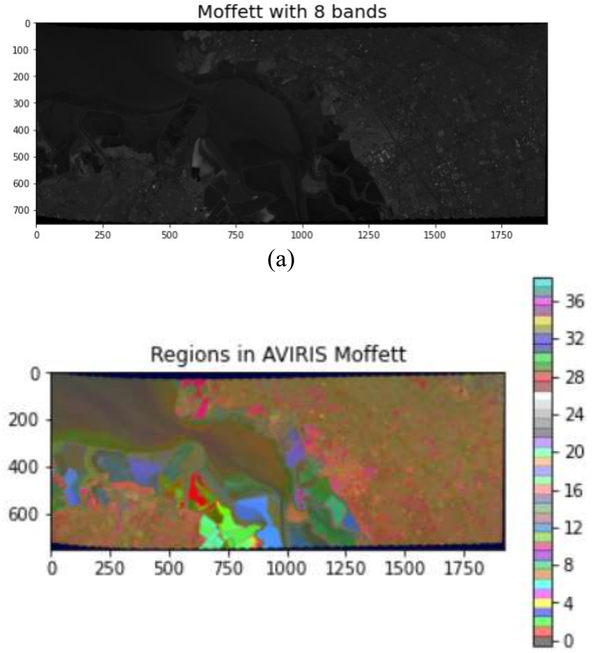
(a) (b)

Fig. 2. Pavia results (a) 8 principal components after PCA. (b) K-means clustering result.



(a) (b)

Fig. 3. Cuprite results (a) 8 principal components after PCA. (b) K-means clustering result.



(a)

(b)

Fig. 4. Moffett results (a) 8 principal components after PCA. (b) K-means clustering result.

VI. Conclusion and Future work

Hyperspectral images contain wide-ranged spectral information that is used to identify and differentiate objects. Classification is the process of assigning objects to classes with homogeneous characteristics. This paper is firsthand examination of: dimensionality reduction via PCA and unsupervised classification via K-means to cluster the three hyperspectral datasets with the goal of finding number of regions based on some similarity in the datasets. Initially, SVM was being used for clustering. However, due to computational expenses, it was dismissed. With K-means, clusters have been formed without any prior knowledge i.e. labels.

One of the limitations of PCA for dimensionality reduction is the loss of spectral information due to selection of principal components based on variance only and not considering the spectral position [15]. Another approach of efficiently clustering data and finding number of regions can be: supervised classification, by breaking the dataset into training and test sets.

Data Set

[1] M Graña, MA Veganzons, B Ayerdi, "GRUPO DE INTELIGENCIA COMPUTACIONAL(GIC)," [Online]. Available: http://www.ehu.eus/ccwintco/index.php/Hyperspectral_Remote_Sensing_Scenes. [Accessed 8 May 2021].

References

[1] T. Subba Reddy and J. Harikiran, "Hyperspectral image classification using support vector machines," IAES Int. J. Artif. Intell., vol. 9, no. 4, pp. 684–690, 2020, doi: 10.11591/ijai.v9.i4.pp684-690.

[2] B. Rasti et al., "Feature Extraction for Hyperspectral Imagery: The Evolution from Shallow to Deep: Overview and Toolbox," IEEE Geosci. Remote Sens. Mag., vol. 8, no. 4, pp. 60–88, 2020, doi: 10.1109/MGRS.2020.2979764.

[3] W. Lv and X. Wang, "Overview of Hyperspectral Image Classification," J. Sensors, vol. 2020, 2020, doi: 10.1155/2020/4817234.

[4] C. R. Addanki, S. A, and V. R. A, "Study of the Clustering Algorithms for Hyper Spectral Remote Sensing Images," J. Hyperspectral Remote Sens., vol. 10, no. 2, p. 117, 2020, doi: 10.29150/jhrs.v10.2.p117-121.

[5] M. J. Khan, H. S. Khan, A. Yousaf, K. Khurshid, and A. Abbas, "Modern Trends in Hyperspectral Image Analysis: A Review," IEEE Access, vol. 6, no. June, pp. 14118–14129, 2018, doi: 10.1109/ACCESS.2018.2812999.

[6] C. Rodarmel and J. Shan, "Principal component analysis for hyperspectral image classification," Surv. L. Inf. Sci., vol. 62, no. 2, pp. 115–122, 2002.

[7] S. Ranjan, D. R. Nayak, K. S. Kumar, R. Dash, and B. Majhi, "Hyperspectral image classification: A k-means clustering based approach," 2017 4th Int. Conf. Adv. Comput. Commun. Syst. ICACCS 2017, pp. 0–6, 2017, doi: 10.1109/ICACCS.2017.8014707.

[8] H. Yadav, A. Candela, and D. Wettergreen, "A Study of Unsupervised Classification Techniques for Hyperspectral Datasets," IGARSS 2019 - 2019 IEEE Int. Geosci. Remote Sens. Symp., pp. 2993–2996, 2019, doi: 10.1109/igarss.2019.8900501.

[9] S. Sattar, H. A. Khan, and K. Khurshid, "OPTIMIZED CLASS-SEPARABILITY IN HYPERSPECTRAL IMAGES," in 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), 2016, pp. 2711–2714, doi: 10.1109/IGARSS.2016.7729700.

[10] I. A. Huqqani and K. Khurshid, "Comparative study of supervised classification of urban area hyperspectral satellite imagery," J. Sp. Technol., vol. 4, no. 1, pp. 1–9, 2014.

[11] A. Rehman, "Project : Hyperspectral Image Principal Component Analysis and Classification Submitted to : Submitted by :," no. April, 2020.

[12] F. Masood, I.-H. Qazi, and K. Khurshid, "Saliency-based visualization of hyperspectral satellite images using hierarchical fusion," J. Appl. Remote Sens., vol. 12, no. 04, p. 1, 2018, doi: 10.1117/1.jrs.12.046011.

[13] S. A., "Hyperspectral Image Classification Using Unsupervised Algorithms," Int. J. Adv. Comput. Sci. Appl., vol. 7, no. 4, pp. 198–205, 2016, doi: 10.14569/ijacsa.2016.070425.

[14] K. Kotwal and S. Chaudhuri, "Visualization of hyperspectral images using bilateral filtering," IEEE Trans. Geosci. Remote Sens., vol. 48, no. 5, pp. 2308–2316, 2010, doi: 10.1109/TGRS.2009.2037950.

[15] H. A. Khan, M. M. Khan, K. Khurshid, and J. Chanussot, "Saliency based visualization of hyper-spectral images," Int. Geosci. Remote Sens. Symp., vol. 2015-Novem, pp. 1096–1099, 2015, doi: 10.1109/IGARSS.2015.7325961.