Mineral Identification Using Unsupervised Classification from Hyperspectral Data

Priyanka Gupta and M. Venkatesan

Abstract Hyperspectral imagery is one of the research areas in the field of remote sensing. Hyperspectral sensors record reflectance of object or material or region across the electromagnetic spectrum. Mineral identification is an urban application in the field of remote sensing of Hyperspectral data. Challenges with the hyperspectral data include high dimensionality and size of the hyperspectral data. Principle component analysis (PCA) is used to reduce the dimension of data by band selection approach. Unsupervised classification technique is one of the hot research topics. Due to the unavailability of ground truth data, unsupervised algorithm is used to classify the minerals present in the remotely sensed hyperspectral data. K-means is unsupervised clustering algorithm used to classify the mineral and then further SVM is used to check the classification accuracy. K-means is applied to end member data only. SVM used k-means result as a labelled data and classify another set of dataset.

Keywords Hyperspectral imagery · Unsupervised classification · K-means · PCA · SVM

1 Introduction

Hyperspectral sensors are used to target detection, classification, pattern recognition and discrimination. These sensors collect images of earth surface in the form of narrow, continuous and discrete spectral bands. These spectral bands form a complete, continuous spectral pattern of each pixel. Most of the study [1], using

P. Gupta (✉) · M. Venkatesan (✉)
Department of Computer Science Engineering,
National Institute of Technology, Surathkal, Karnataka, India
e-mail: prigupta9875@gmail.com

M. Venkatesan
e-mail: venkisakthi77@gmail.com

P. Venkata Krishna and M. S. Obaidat (eds.), *Emerging Research in Data Engineering Systems and Computer Communications*, Advances in Intelligent Systems and Computing 1054, https://doi.org/10.1007/978-981-15-0135-7_25

hyperspectral data for geological applications, have so far addressed in the different regions of climates. A mineral or combinations of minerals are source of material, known by a combination of different minerals that is in many different forms like solid, organic and inorganic. Each pixel contains a mixture of different spectra due to the multiple components available in the surface that form the ground surface. This complexity results in incorrect identification and/or misclassification of surface materials. Therefore, the classification of materials and minerals from different areas of earth's surface is one of the most important research topics in remote sensing of hyperspectral data.

Hyperspectral image classification can be of three types—supervised, semi-supervised and unsupervised classification based on ground truth data availability. The semi-supervised classification [2] is where we used some labelled data and based on that classifying the unlabelled data. However, in unsupervised classifier [3], a remote-sensing image is divided into a number of groups of similar characteristic of the image values and then classified into classes, without any knowledge of ground truth data. Two unsupervised classification algorithms—k-means and its variant, and the iterative self-organizing data analysis (ISODATA) technique—are the most commonly used classifiers. Both give the same set of clusters when a number of clusters are same. In k-means number of cluster knowledge is priori while in ISODATA no need of prior knowledge of number of clusters. In this work, we focus on some unsupervised classification technique and combine with some supervised technique. Here, we used the k-means and SVM classifier with sigmoid kernels to get better classification accuracy.

Hyperspectral remote-sensing datasets are represented as a 3D data cube with spatial and spectral information such that X–Y plane contains spatial information and Z-direction contains spectral information. The hyperspectral datasets have more than 200 narrow and contiguous wavelength bands at bandwidths of about 5–10 nm. Dataset used for this study is obtained from EO-1 Hyperion satellite which does not has the ground truth data.

2 Background

Mineral mapping is used to map different types of minerals with their contents and characteristics. It is one of the important applications in high resolution of remote sensing data by hyperspectral technique. So in the application of remote sensing technology and hyperspectral technology, the analysis of rocks and minerals has superiority to traditional ways. Many research have been presented by analysing mineral spectra of visible and near-infrared (VNIR) bands which give a promising identification model of minerals with acceptable accuracy to acquire the mineral category and content in rock images taken from hyperspectral sensor. Satellite and airborne images are used to classify crops, examine their health, finding disease, etc.

Classification of hyperspectral image are based on both spectral and spatial features. In reference to [4], their proposed work is based on low spatial resolution because spatial features can be changed within metres that affect different factors of the images, such as imperfect imaging and atmospheric scattering while detecting reflectance. Other factors which degrade image quality are sensor noise and secondary illumination effect, spatial resolutions helps to remove these effects to improve in quality of hyperspectral images.

In earlier study, many of the works are done in the field of remote sensing and mineral exploration in different areas. According to [5], work is done on spectral analysis and mapping of different minerals in part of Latehar and Gumla District, Jharkhand. In this study, used EO-1 Hyperion data for AL + OH mineral from rocks first compensate the atmospherical effect from data and MNF transformation is used to reduce the data noise from it. To find out the end member as the target member apply singular angular mapping and matched filtering in it.

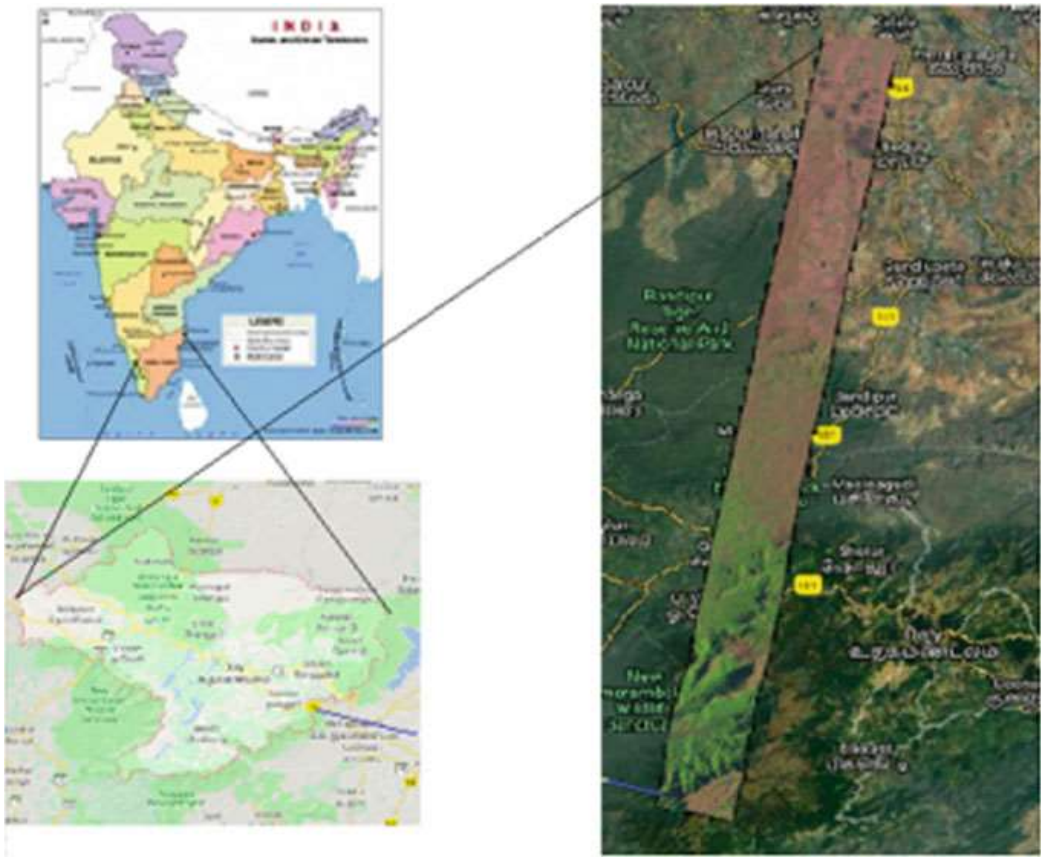

Fig. 1 Satellite view of Nilgiris hill

3 Study Area and Data

Study area [6] is located in the state of Tamil Nadu, southern India at latitude 11 08′ to 11 37′ N and longitude of 76 27′ E to 77 4′ E shown in Fig. 1. Minerals present in Nilgiris district of Tamil Nadu are iron ore, bauxite and clay utilize in different industries. There is an excellence of different types of minerals distributed in different regions with different compositions. Iron ore and bauxite are found near Kotagiri. Magnetize is found close to Thengumaranada in Moyar valley and Quartz is in Devala and Clay in the constitution of China clay arises in Cherambadi of Nilgiris district.

Hyperion EO-1 is a US satellite which is used to collect the data from space. This type of data has 242 narrow continuous spectral bands with spectral variety between 0.4 and 2.5 m at 10 nm interval calibrated with 16 bit resolution and 30 m spatial resolution.

4 Methodology

In this section, we described a pixel-based classification technique in unsupervised way.

4.1 Data Pre-processing

Pre-processing in high dimensional data is one of the important challenges. Pre-processing makes the data more accurate and noise-free; so that, it can give better result. Pre-processing will remove all the dead pixels, water band, noisy pixels, etc. In hyperspectral data, pre-processing will remove all bad band, noisy band and zero band instead of pixels.

Zero Band Removal
The bands which do not have any pixel information in hyperspectral data is called zero band. By using ENVI software (used for visualizing a geological data), we visualize that some set of bands are zero band which we need to remove it. In EO-1 Hyperion dataset's, zero bands are listed in Table 1.

Table 1 List of zero band

S. No.	Zero bands	Reason
1.	1–7	Zero bands
2.	58–78	Overlap region
3.	120–132, 165–182, 218–224	Water vapour absorption
4.	184–186, 225–242	Bad bands

Destriping of Band

Vertical stripe may occur in the region where brightness of pixel varies relatives to nearby pixels. These stripes make the image unclear and contain pixels with wrong information that will give a negative impact on further processing algorithms. Using local destriping algorithm, we can remove this type of strips to some extent. Used equation by the algorithm is

$$\sum_{j=1}^{n}[(x_i - 1, j, k) + (x_i + 1, j, k)]/2n \tag{1}$$

Atmospheric Corrections

The reflected solar energy travels through the atmosphere. Based on the amount of atmospheric reflection, types of particles and gases available, atmospheric absorption and atmospheric scattering, light interacts with atmosphere and materials and reflected energy store in the form of spectrum. Atmospheric correction is compulsory to remove all these unwanted effects. FLAASH stands for fast line-of-sight atmospheric analysis of hypercube and able to process wavelengths in VNIR and SWIR region up to 3 μm. In this study, we used ENVI for atmospheric correction. FLAASH will also be able to remove adjacency consequence, cirrus and opaque cloud map and also compute a scene-average visibility. This removes all water vapour windows from the images. After this step, all the bands present in the dataset are noise-free. FLAASH will also be able to remove adjacency consequence, cirrus and opaque cloud map and also compute a scene-average visibility. This removes all water vapour windows from the images. After this step, all the bands present in the dataset are noise-free.

4.2 Dimensionality Reduction

Principle component analysis (PCA) is a dimensionality reduction technique. It is used as feature extraction, and in hyperspectral data, feature extraction can be done by band selection. This process is defined as reducing the number of random variables under consideration, by obtaining a set of principal component, i.e. selection of band. In high dimensional data, feature extraction can minimize execution time for hyperspectral data. PCA is based on eigenvalue decomposition of covariance matrix. Let us consider hyperspectral image is of $M \times N \times B$ size. Pixel vector is calculated using all bands as in stack as shown in Fig. 2. In [7], show pixel vector in hyperspectral images.

$$X_i = [x_1, x_2, x_3, \ldots, x_n]$$

where B is number of band and M and N are number of rows and columns, respectively, i = 1, 2, 3, …, M_1 and

$$M_1 = M \times N$$

Mean will be calculated by

$$m = \frac{1}{M_1}\sum_{i=1}^{M_1} X_i = ([x_1, x_2, x_3, \ldots, x_n])^{\mathrm{T}}$$

Covariance matrix will be calculated by

$$C_x = \frac{1}{M_1}\sum_{i=1}^{M_1}(X_i - M)(X_i - M)^{\mathrm{T}}$$

For eigenvalue decomposition of the covariance matrix

$$C_x = \mathrm{ADA}^{\mathrm{T}}$$

where $D = \mathrm{diag}(\lambda_1, \lambda_2, \lambda_3, \ldots, \lambda_N)$ is the diagonal matrix composed of the eigenvalues $\lambda_1, \lambda_2, \lambda_3, \ldots, \lambda_N$ of the covariance matrix C_x and A is the orthonormal matrix composed of the corresponding N dimension eigenvectors ak(k = 1, 2, …, n) as follows:

$$A = (a_1, a_2, \ldots, a_n)$$

The linear transformation is

$$Y_i = A^{\mathrm{T}} X_i (i = 1, 2, 3, \ldots, M_1)$$

is the PCA pixel vector, and all these pixel vectors form the PCA bands of the original images.

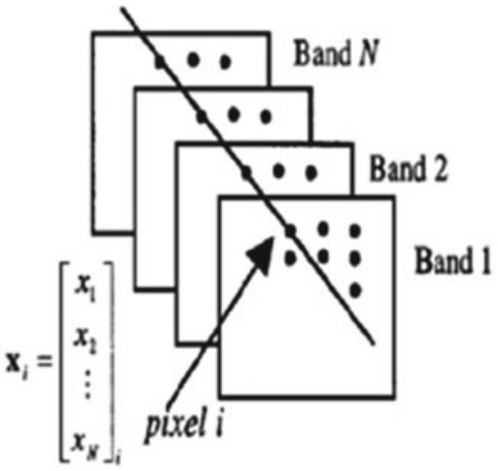

Fig. 2 Pixel vector

4.3 Classification Approach

Unsupervised classification can be defined as the identification of classes by grouping the pixels of similar type within one class. Based only on their statistics without using previous knowledge about the spectral classes presented in the image, pixels are clustered into classes. K-means is one of the commonly used methods in unsupervised classification. But we need prior knowledge of number of classes is prerequisite in K-means. To estimate the number of cluster, we used elbow method.

Elbow Method

To find out the *k* (number of cluster) value, use elbow method. According to [8], it examines the percentage of variance as a function of cluster *k*. It runs k-means on the dataset for the range of value *k,* e.g. 1–10 and for each value of *k,* it calculates the sum of squared error (SSE). And then, it plots a graph of SSE for each value of *k*. After that observe the graph and check that at which value of k graph go flatten that would be an optimal number of cluster. In Fig. 3, red-dotted circle shows that after 4, the graph start flatten and after 6, it becomes flat. So, the optimal number of classes possible is six, i.e. *k* will be equal to six.

K-means Clustering

After finding the number of classes, next, we run k-means on the extracted end member. The main idea behind k-means is to initialize one value to each cluster which is called centroid or mean. These centroid values are assigned to each clusters and then pick each pixel value and assign it to those cluster which nearest to centroid. Next task is to recompute the centroid for each cluster and repeat the process until convergence criteria are met.

We applied k-means only on extracted end members which give high probability of correct classification. It classifies the image into four classes such that each index of the image is corresponding to one of the class. Because we do not have ground

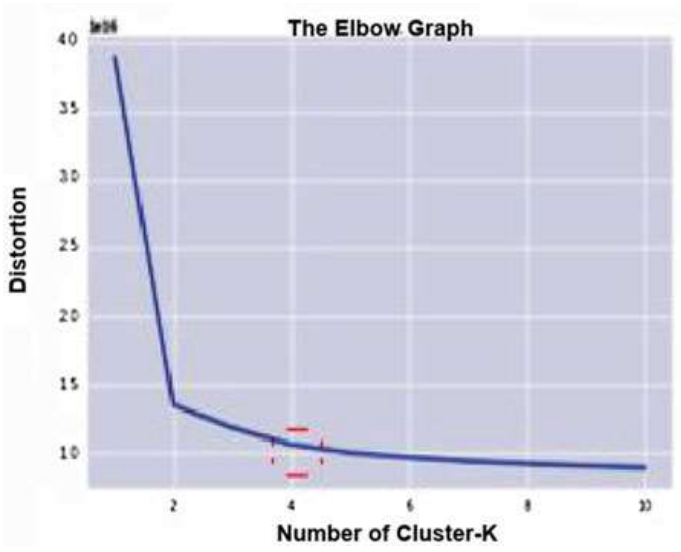

Fig. 3 Elbow graph

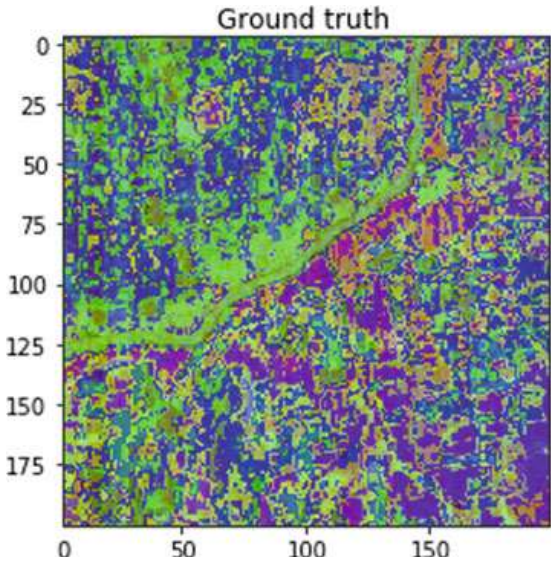

Fig. 4 Ground truth generated by k-means

truth data we consider this as a ground truth data. Figure 4 shows the obtained ground truth data. Because ground truth data is not available to check the accuracy of clustering method is challenging task. Still, there are some methods used to check that how better is clustering result. Davies-Bouldin index is one of those methods that can be used to evaluate the model. Where, a lower DB index relates to a model with better separation between the clusters and vice versa. In this k-means algorithm DB index score is equal to 0.316.

SVM Classifier

In this work, support vector machine considers for multiclass problem of hyperspectral imagery. SVM is supervised technique, so k-means result is considered as a ground truth labelling. According to [9], SVMs perform better than other classification techniques and also in pattern recognition and provide higher classification accuracy. Furthermore, SVMs also give good accuracies even in the presence of heterogeneous classes for which only few training samples are available. In this, different kernels can be used based on dataset and set of problems. Here used kernel in tangent-based function called as sigmoid function.

5 Experiment and Result

In proposed method, hyperspectral image is of size $M \times \leftarrow N \times B$ size. Where B is the number of bands in image and $M \times N$ is size of each band. Total number of bands in dataset is 224 after removal of all bad bands and noise correction left band are 138 out of 224. Next PCA is applied on left bands. And out of 138 bands, 12 bands are selected as an end member. On these end members, we used k-means clustering which classifies the image into six classes. Now we train our original sample with these labelled data (as shown in Fig. 4), using SVM classifier.

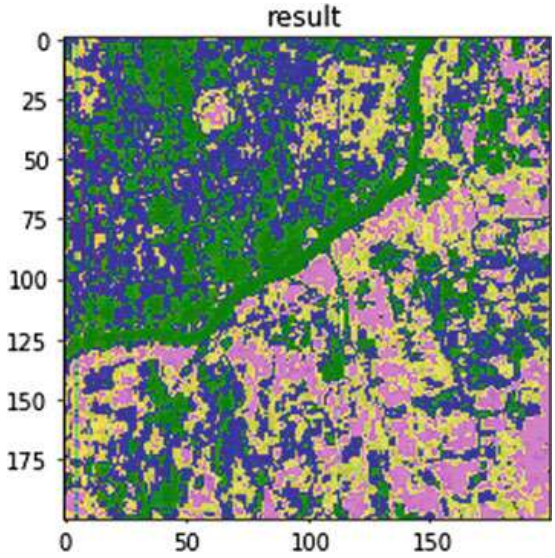

Fig. 5 Predicted result using SVM

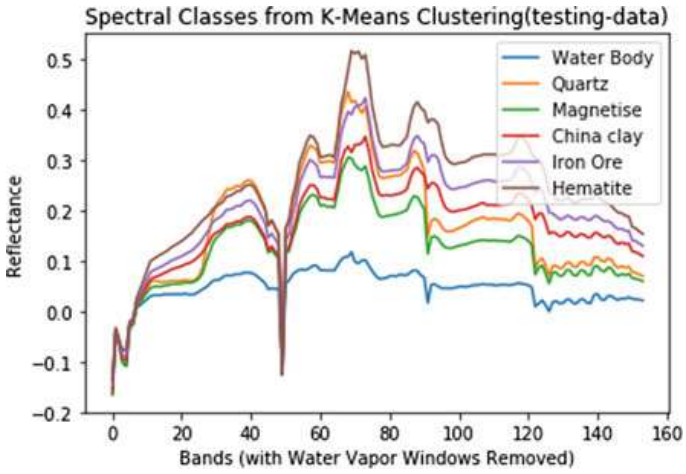

Fig. 6 Spectra profiles of minerals

We used some part of dataset as a test data on which we predict result using SVMs as shown in Fig. 5. This is able to classify into four classes successfully and SVM gives accuracy score as 76.03%. Spectra profile of these six classes is shown in Fig. 6. In which, green spectra profile shows water body and rest of the types of minerals. The range of reflectance is in between 0.4 and 2.5, therefore in graph, spectra profile exists from 0.3 to up to 3 μm.

6 Conclusion

Mineral classification from the hyperspectral data is unsupervised classification because labelled data is not available. K-means is used here to classify the data into desired classes. Hyperion data with 242 bands used for classifying minerals.

After removing all noise from the data, band selection performed using PCA. Then, K-means applied on selected band and got the labelled data which is further used for SVM classifier. SVM classifier gives approximately 76.03% accuracy. This classifier also depends on the dataset as well as on different parameters used in SVMs. For the future work, we can work on classifier to improve the accuracy.

References

1. Zhu, X.X., Tuia, D., Mou, L., Xia, G.-S., Zhang, L., Xu, F., Fraundorfer, F.: Deep learning in remote sensing: a comprehensive review and list of resources. IEEE Geosci. Remote Sens. Mag. **5**(4), 8–36 (2017)
2. Sawant, S.S., Prabukumar, M.: Semi-supervised techniques based hyper-spectral image classification: a survey. In: Power and Advanced Computing Technologies (i-PACT), 2017 Innovations in IEEE, pp. 1–8 (2017)
3. Mou, L., Ghamisi, P., Zhu, X.X.: Unsupervised spectral–spatial feature learning via deep residual conv–deconv network for hyperspectral image classification. IEEE Trans. Geosci. Remote Sens. **56**(1), 391–406 (2018)
4. Villa, A., Chanussot, J., Benediktsson, J.A., Jutten, C., Dambreville, R.: Unsupervised methods for the classification of hyperspectral images with low spatial resolution. Pattern Recogn. **46**(6), 1556–1568 (2013)
5. Satpathy, R., Singh, V.K., Parveen, R., Jeyaseelan, A.T.: Spectral analysis of hyperion data for mapping the spatial variation of AL + OH minerals in a part of Latehar & Gumla district, Jharkhand. J. Geogr. Inf. Syst. **2**(4), 210 (2010)
6. Vigneshkumar, M., Yarakkula, K.: Nontronite mineral identification in Nilgiri hills of Tamil Nadu using hyperspectral remote sensing. In: IOP Conference Series: Materials Science and Engineering, vol. 263, no. 3, p. 032001. IOP Publishing (2017)
7. Ranjan, S., Nayak, D., Satish Kumar, K., Dash, R., Majhi, B.: Hyperspectral Image Classification: A k-means Clustering Based Approach, pp. 1–7 (2017)
8. Kingrani, S.K., Levene, M., Zhang, D.: Estimating the number of clusters using diversity. Artif. Intell. Res. **7**(1), 15 (2017)
9. Moughal, T.: Hyperspectral image classification using support vector machine. In: J. Phys.: Conference Series, vol. 439, no. 1, p. 012042. IOP Publishing, (2013)