Land cover clustering and classification of satellite images

Vaishnavi Kharat[1*], *Sanyukta* Khatdeo[1], *Harshada* Kothe[1], *Rutuja* Kshirsagar[1], *Mrudul* Dixit[1], *M Selva* Balan[2]

[1]Electronics & Telecommunication, MKSSS's Cummins College of Engineering for Women, Pune, India.
[2]Hydraulic Instrumentation, Central Water and Power Research Station (CWPRS), Pune, India

Abstract. Land cover classification refers to the process of using remote sensing data to categorize different types of land cover like vegetation, water bodies and soil. This is helpful for gaining key information about the surface of the Earth and for the future interactions between human activities and the environment. These predicted interactions lead to the development of sustainable land use practices along with the protection of natural resources. This paper deals with classifying the land cover using unsupervised and supervised methods. The unsupervised method includes land cover detection using a K-means clustering algorithm and the supervised classification is done using random forest classifier. The evaluation parameter values are calculated and compared for the input and output images.

Keywords—classification, remote sensing, multiband satellite imagery, rescaling, segmentation

1 Introduction

Land cover refers to the features that cover the Earth's surface, consisting of natural and human-made features. It includes vegetation, water bodies, bare soil and other land use types. Classifying this land cover is an important step in order to understand the biophysical system of Earth. Classified land cover helps to monitor the health conditions of ecosystems, as well as to estimate the availability and distribution of natural resources.

The land cover classifier is also used as a tool to study the impact of climatic changes on ecosystems and to assess the vulnerability of areas to disasters like floods, landslides and forest fires.

The objective of this research is to classify land cover from satellite multiband images consisting of 12 bands. Satellite data is classified into three main areas, namely land, water and vegetation.

*corresponding author: vaishnavi.kharat@cumminscollege.in

The objective of this research is to provide CWPRS, Pune with satellite multiband images with land cover classification, in order to help achieve future requirements for the development of optimal and secure water resources projects in the country.

2 Methodology

2.1 Unsupervised Classification using K-means clustering

The clustering algorithm is performed on a set of remote sensing images, by importing several modules including 'matplotlib', 'gdal', 'numpy', 'pandas', and 'imageio', used to read image, sample a subset of pixels, fit a k-means model to the samples, and display the results. It outputs a pair plot of the band values, a pair plots with colored clusters, and a map of the clustered data.

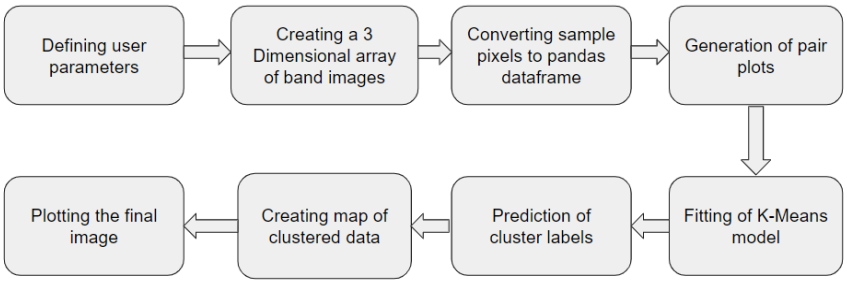

Fig. 1. Block diagram of K means Algorithm

2.1.1 Defining user parameters:

The user-defined parameters must be adjusted to fit the specific images and parameters that are worked with.
'image_folder_name' : the path to the folder where the image files are located.
'image_format' : the file format of the image files (tiff).
'band_names' : a list of the names of the bands that are included in the image files. The names should match the naming convention used in the file names.
'Nsamples' : the number of randomly sampled pixels to use in training the k-means model. A large value will result in more accurate clustering, but will also increase the time required to run the script.
'NUMBER_OF_CLUSTERS' : the number of clusters to use in the k-means model. This should be chosen based on the user's prior knowledge of the data and the desired level of granularity in the clustering.
'clour_map' : the color map is used in the output image.

2.1.2 Creating 3D array of band images:

The band images in the folder are imported into a dictionary and a 3D array of zeros is created with dimensions equal to any of the input images in the folder. The number of layers is equal to the number of bands. The pixel values of each band are stored on it.

2.1.3 Converting sample pixels to pandas dataframe:

A random subset of images is chosen and is then converted to pandas dataframe. In this dataframe, columns correspond to the band names while the rows correspond to each pixel sample.

2.1.4 Generation of pair plots:

Generation of pair plots is done to visualize the correlation (Bivariate distribution) between the bands. Each band is plotted against another band. This is done using the seaborn library. The parameters are stored in a dictionary and then corresponding histograms are plotted.

2.1.5 Fitting of K means model:

A K means a clustering model is fitted over the data. The number of clusters is assigned to the user parameters. A cluster label is assigned to each sample.

2.1.6 Prediction of cluster labels:

The fitted model is used to predict the cluster assignments for the samples selected. The resulting cluster assignments are then used to color code the pairwise scatterplots of the pixel values of different bands

2.1.7 Creating map of clustered data:

Cluster assignments are stored in a temporal variable. A new array is created with the same dimensions as the initial array. The clustered labels for each row are stored in the new array.

2.1.8 Plotting the final image:

Image clustered array is plotted and a colormap is applied to cluster labels to create color coded representation. A final clustered image is hence displayed.

2.2 Supervised Classification using Segmentation and Random Forest Classifier

The Land cover classification of satellite images using segmentation and random forest is a common approach in remote sensing and image processing. The process involves dividing the satellite image into segments, which are regions with similar spectral properties, and then using a random forest classifier to classify the land cover within each segment.

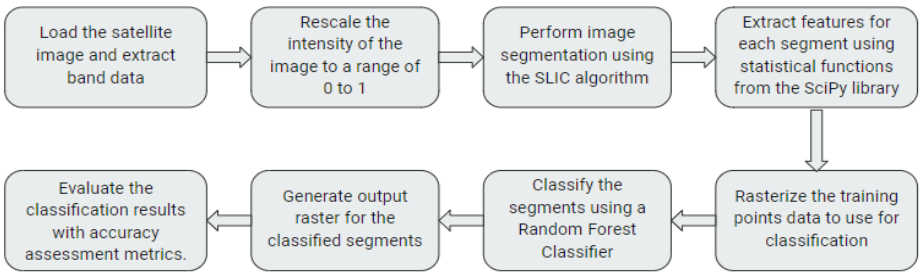

Fig. 2. Classification using Segmentation and Random Forest

2.2.1 Loading the satellite image and extract band data:

The GeoTIFF file located at the specified file path is opened using the GDAL library, then, the raster bands of the GeoTIFF file are read using a for loop and stores the band data in a

NumPy array. Finally, the bands are stacked in the NumPy array using the dstack() function from NumPy. The purpose is to prepare the data for image segmentation using the SLIC algorithm from the skimage library.

2.2.2 Rescaling the intensity of the image to a range of 0 to 1:

The rescaling is performed to ensure that the pixel intensities fall within a certain range, i.e. 0 and 1. The rescaling is performed by stretching or shrinking the range of the pixel values in the input image to fit within the specified range. The rescaled image can be used as input for further image processing steps.

2.2.3 Performing image segmentation using the SLIC algorithm:

The rescaling is performed to ensure that the pixel intensities fall within a certain range, i.e. 0 and 1. The rescaling is performed by stretching or shrinking the range of the pixel values in the input image to fit within the specified range. The rescaled image can be used as input for further image processing steps.

2.2.4 Extracting features for each segment:

The SLIC algorithm works by clustering pixels that are similar to each other in color and texture, and then grouping those clusters together to form segments. Two parameters are specified for the SLIC algorithm: “n_segments” and “compactness”. “n_segments” controls the number of segments that the algorithm will produce. “compactness” controls the balance between color proximity and spatial proximity when grouping pixels into clusters. A smaller value for “compactness” results in clusters that are more spatially compact, while a larger value results in clusters that are more color homogenous (here, set it to 0.1).

2.2.5 Rasterizing the training points data to use for classification:

A function “segment_features” is defined that takes a NumPy array of pixels for a single image segment as its argument. For each band, the function calculates the minimum, maximum, mean, variance, skewness, and kurtosis, it then combines the statistical features for all bands into a single list and returns it as the output of the function, which can then be used as input to ML algorithms for classification. The resulting SLIC segmentation map is saved as a new GeoTIFF file.

2.2.6 Classifying the segments using a Random Forest Classifier:

A new raster layer is created, the training data points are rasterized from the shapefile onto the new raster layer. It retrieves the rasterized data as a NumPy array and prints some basic statistics, such as the minimum, maximum, and mean values, and reads the rasterized training data. The unique land cover classes are identified from the ground truth data by removing the background value (0), a dictionary is then created called "segments_per_class" that associates each land cover class with the set of segment IDs that correspond to that class in the ground truth data.

2.2.7 Generating output raster for the classified segments:

Supervised classification is performed using the Random Forest algorithm. It uses a raster image as input (naip_fn) and a vector shapefile containing training data (train_fn). The training data is obtained by rasterizing the vector shapefile. The training objects are obtained by selecting the statistical features for each segment that belongs to a land cover type, and the training labels are simply the class IDs. Finally, the Random Forest classifier is trained using the fit method.

2.2.8 Evaluate the classification results with accuracy assessment metrics:

Finally, evaluate the performance of a classification model by computing the confusion matrix and other performance metrics such as accuracy, precision, recall and F1-score to evaluate the performance of a classification model by showing how many of the actual values were correctly predicted and how many were incorrectly predicted.

3 Results

3.1 Unsupervised Classification using K-means clustering

Fig. 3. Input multiband satellite image in TIF format

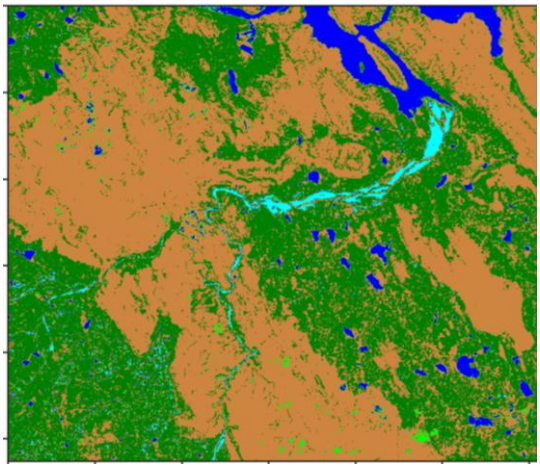

Fig. 4. Output of K-means clustering
(Number of clusters=5, =Water, =River tributary, =barren Land, =Vegetation, =Grassland)

3.2 Supervised Classification using Segmentation and Random Forest Classifier

Fig. 5. Cloudless multiband satellite image taken as input

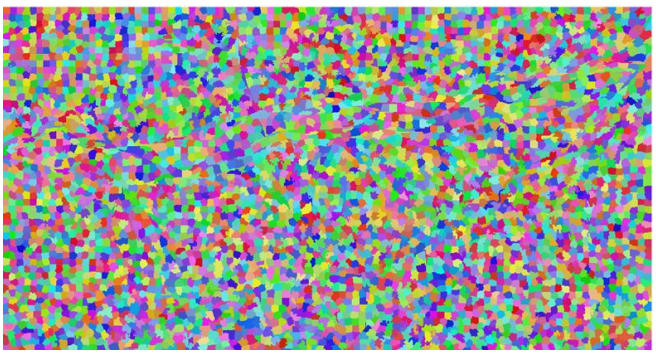

Fig. 6. Output of segmentation of cloud free multiband satellite image using SLIC algorithm when n_segments=5000 and compactness=0.1

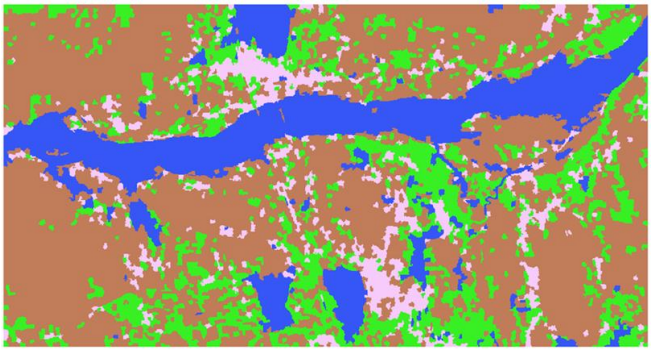

Fig. 7. Output of classification using Random Forest classifier

=Water =Vegetation =Land =Urban

4 Evaluation parameters

4.1 Silhouette score: It is used to determine a data point's position within its own cluster in relation to others. Scores for silhouette range in between -1 and 1.

4.2 Davies Bouldin Index (DBI): It states the similarity between clusters by computing the ratio of total of distances between the centroids of each couple of clusters to distance between centroids of two closest clusters. A lower DBI is used to indicate better performance.

4.3 Calinski-Harabasz Index (CHI): This is the ratio between cluster dispersion to within cluster dispersion. CHI should have a high value for better performance.

4.4 Confusion Matrix: This grid summarizes the execution of a classification model. Each cell reflects a number of samples belonging to the true class and have been predicted as belonging to the predicted class. The diagonal represents correctly classified samples.

4.5 Accuracy: The percentage of samples that are properly categorized of all samples.

4.6 Precision: It is the proportion of correctly classified positive samples out of all samples predicted as positive.

4.7 Recall: It is the proportion of correctly classified positive samples out of all samples that are actually positive.

4.8 F1 score: It is the weighted average of recall and precision.

Table 1. Evaluation parameters for output image using K means clustering algorithm.

Parameter	Value
Silhouette Score	0.46086
DBI	0.6831
CHI	11167.0509

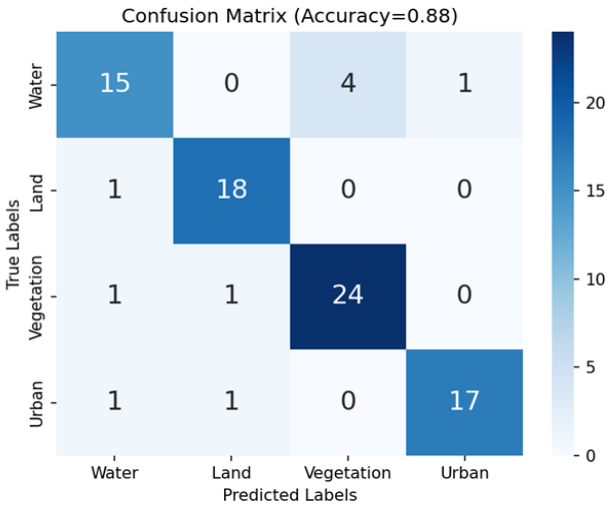

Table 2. Evaluation parameters for output image after supervised classification.

Parameter	Value
Accuracy	0.88091
Recall	0.88095
F1 score	0.87974

5 Application

Satellite image classification of land cover is a useful tool with many uses in various industries. This method has many important applications, including environmental monitoring, which makes use of it to monitor changes in vegetation cover, deforestation, land use, and land degradation. Effective environmental planning and management require this knowledge. Identifying crop types, evaluating crop health, and estimating yields can all be done using land cover categorization in agriculture, allowing for efficient crop management and planning. Urban planning, disaster management, and resource allocation can all be aided by using land cover classification to identify urban areas, roads, buildings, and other infrastructure. Furthermore, land cover classification helps managers of natural resources to identify forests, water bodies and wetlands. Monitoring changes in the Earth's surface over time and evaluating the effects of natural catastrophes are important for emergency response and studies on climate change. Overall, the ability to identify land cover from satellite photos is a crucial tool for a number of tasks, such as monitoring the environment, farming, urban planning, managing natural resources, studying climate change, and responding to emergencies.

6 Conclusion

Unsupervised learning is a method that uses K-means clustering. Training data with labels is not necessary. Each group or cluster represents a different type of land cover, and the algorithm organizes pixels in the image based on how similar they are. By minimizing the

sum of squared distances between each pixel and its designated cluster centroid, the algorithm determines the ideal number of clusters. When analyzing huge datasets, K-means clustering can be a quick and effective way for classifying land cover.
Using segmentation and random forest, land cover classification of satellite images is a supervised classification technique for mapping the distribution of land cover categories over vast areas. This method combines the advantages of random forest and segmentation.

References

1. M. Abedini, *Clustering Approach on Land Use Land Cover Classification of Landsat TM Over Ulu Kinta Catchment*, (2012)
2. Yanqun Pan 1, Simon Bélanger and Yannick Huot "*Evaluation of Atmospheric Correction Algorithms over Lakes for High-Resolution Multispectral Imagery: Implications of Adjacency Effect*", Remote Sens. 2022, 14, 2979, (2022)
3. Akhtar Alam. M. Sultan Bhat. M. Maheen- "*Using Landsat satellite data for assessing the land use and land cover change in Kashmir valley*" GeoJournal (2020) 85:1529–1543, (2019)
4. Ksenia S. Yankovich, Elena P. Yankovich and Nikolay V. Baranovskiy- "*Classification of Vegetation to Estimate Forest Fire Danger Using Landsat 8 Images:Case Study*", Hindawi, Mathematical Problems in Engineering,Volume 2019, Article ID 6296417 (2019)
5. E. Jargaldalai, D. Amarsaikhan, M.-E. Altangerel, and B. Tovuusuren, "*Object-Based Classification of Land Cover types using Lidar and Satellite Images*," ResearchGate, (2022)
6. A. Smith, "*Image segmentation scale parameter optimization and land cover classification using the Random Forest algorithm,*" Journal of Spatial Science, vol. 55, no. 1, pp. 69–79, doi: 10.1080/14498596.2010.487851 (2010)