Relational Features of Remote Sensing Image Classification using Effective K-Means Clustering

T. Balaji ¹, M. Sumathi ²

¹ Assistant Professor, Dept. of Computer Science, Govt. Arts College, Melur, India
² Associate Professor, Dept. of Computer Science, Sri Meenakshi Govt. College for Women, Madurai, India

ABSTRACT

The feature based classification of remotely sensed image is used to assign corresponding levels with respect to groups with homogeneous characteristics, with the aim of discriminating multiple objects from each other within the image. Every level of an image is called class. This will be executed on the basis of spectral or spectrally defined features such as density, texture and many other things in the feature space. This paper focuses remote sensing image classification of color feature based using k-means clustering method. K-means is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters. The main idea is to define k centroids, one for each cluster. These centroids should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and we need to re-calculate k new centroids of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid. A loop has been generated. As a result of this loop we may notice that the k centroids change their location step by step until no more changes are done. Here we introduce several widely used algorithms that consolidate data by clustering or grouping and then present a suitable method is remote sense application based k-means cluster algorithm. It is possible to reduce the computational cost and gives a high discriminative power of regions present in the image.

Keywords: Classification, Clustering, Unsupervised, Segmentation and Partitioning

1 INTRODUCTION

Remote sensing image classification refers to the task of extracting information classes from a multiband raster image. The resulting raster from image classification can be used to create thematic maps. Depending on the interaction between the analyst and the computer during classification, there are two types of maps: supervised and unsupervised. The objective of image classification is to identify and portray, as a unique gray level (or color), the features occurring in an image in terms of the object or type of land cover these features actually represent on the ground. A broad group of digital image processing techniques of remote sensing application explains the image classification techniques are most generally applied to the spectral data of a single-date image or to the varying spectral data of a series of multidimensional images. Clustering involves a set of point into non-overlapping groups or points; where points in a cluster are more similar to one another than points in other clusters. The term more similar, when applied to clustered points, usually means closer by some measure of proximity. When a dataset is clustered, every point is assigned to some cluster and every cluster can be characterized by a single reference point, usually take an average of the points in the cluster. Any particular division of all points in a dataset into clusters is called a partitioning. Clustered data require considerably less storage space and can be manipulated more quickly than the original data. The value of particular clustering method will depend on how relatively the reference points represent the data as well as how fast the program runs.

A general classification procedure of remotely sensed image is shown in the following fig. 1.

Fig. 1. General Procedure for Classification
2 CLUSTERING METHOD

Clustering is a grouping of data with similar characteristics. This "similarity" in a given set may vary according to data, because clustering is used in various fields such as numerical taxonomy, morphometric, systematics, etc. Thus, a clustering algorithm that fits the numerical measure of optimization in a data may not optimize another set of data (for example, depending on the units selected). There are many algorithms to solve a clustering problem. The algorithms used in our applet concentrate on "joining", "splitting", and "switching" search methods (also called bottom up, top down, and interchange, respectively). The clustering method used is a three-dimensional model for demonstration purposes. However the algorithms can calculate clusters in n-dimensions. We are planning to an algorithm based on principal component analysis, projecting the n-dimensions in a three dimensional space. There are many methods of clustering developed for wide variety of purposes. Clustering algorithms used for unsupervised classification of remote sensing data vary according to the efficiency with which clustering takes place. Clustering is divided into hierarchical clustering and non-hierarchical clustering as mentioned as follows. The distances to evaluate the similarity are selected from the following methods.

2.1 Hierarchical Clustering Method

- Nearest Neighborhood Method – Nearest neighbor with minimum distance will from a new merged cluster.
- Furthest Neighborhood Method – Furthest neighbor with minimum distance will from a new merged cluster.
- Centroid Method – Distance between the gravity centers of two clusters is evaluated for merging a new merged cluster.
- Group Average Method – Root mean square distance between all pairs of data within two different clusters are used for clustering.
- Ward Method – Root mean square distance between the gravity center and each member is minimized.

2.2 Non-hierarchical Clustering Method

At the initial stage, an arbitrary number of clusters should be temporarily chosen. The members belonging to each cluster will be checked by selected parameters or distance and relocated into the more appropriate clusters with higher separability. The k-mean method is a best example of non-hierarchical clustering. This method is composed of the following procedure:

- All members are relocated into the closest clusters by computing the distance between the member and the clusters.
- The center of gravity of all clusters is recalculated and the above procedure is repeated until convergence.
- If the number of clusters is within a certain specified number and the distances between the clusters meet a prescribed threshold, the clustering is considered complete.

3 K-MEANS CLUSTERING METHOD

The k-means clustering algorithm is faster than the standard version and extends the size of the datasets that can be clustered. It differs from the standard version of the cluster algorithm in how the initial reference points are chosen and how data points are selected for the updating process. In the standard algorithm the initial reference points are chosen more or less arbitrarily. The proposed k-means cluster algorithm reference points are chosen at random sampling from the whole population of data points. If the sample is sufficiently large, the distribution of these initial reference points should reflect the distribution of points in the entire set. Another difference between the standard and the new k-means clustering algorithms is the way the data points are treated as random sampling. The each complete iteration, the standard algorithm examines all the data points in sequence. In contrast, the proposed k-means clustering algorithm examines only a random sample of data points. If the data set is very large and the sample is representative of the dataset, the algorithm should coverage much more quickly than the standard algorithm that examines every point in sequence of updating the centroids during the initial partitioning, when the data point are first assigned to clusters.

Clustering algorithm uses an interchange (or switching) method to partition a graph into clusters. An initial partition is given, and new partitions are obtained by switching an object from one cluster to another. This method randomly picks k points initially, where each point stands for a cluster to be made. A set of points is taken from the graph, and each point is added to the closest cluster. The "closeness" to the cluster is determined by calculating the distance between a point and the centroid of a cluster. Then each point is visited to recalculate the distance to the updated clusters. If the closest cluster of the point is not the one it currently belongs, the point will switch to the new cluster. When switching occurs, centroids of both modified clusters have to be recalculated. This procedure is repeated until no more switching takes place. The K means clustering algorithm in the following logical representation:

- Determine the centroid coordinate (Random assignment).
- Determine the distance of each Object pixel to the centroids.
- Group the object based on minimum distance with the centroid.

4 RELATIVE ERROR MEASURING FACTOR
In fact, the clustering algorithm adopts mathematical method of updating centroids during the initial partitioning, when the data points are first assigned to clusters.

4.1 Error Measurable Factor

From a mathematical perspective, random sampling represents a return to original concept of the algorithm as a method of clustering data over a sequence space. The error measure EMF for each region Ri is given by

$$\text{EMF}_i = \int_{R_i} \sum_{x} q(x) \| x - z_i \|^2 \, dx$$

Where $q(x)$ is the probability density function, a sequence function defined over the space, and the total error measure EMF is given by the sum of the EMF’s. In this mathematical concept of the algorithm, very large set of discrete data points can be thought of as a large sample and is a good estimate of the continuous probability density $q(x)$. It becomes a random sample of the dataset can also be a good estimate of $q(x)$. Such a sample yields a representative set of cluster centroids and a responsible estimate of the error measure without using all the points in the original dataset.

These modifications to the standard algorithm greatly accelerate the clustering process. Since both the reference points and the data points for the updates are chosen by random sampling, more reference points will be found in the densest regions of the dataset and the reference points will be updated by data points in the most critical regions. This clustering algorithm is about ten times faster than the standard algorithm. The computation time can be further reduced by making the individual steps in the algorithm more efficient. A substantial fraction of the computation time required by any of these clustering algorithms is typically spent in finding the reference point closest to a particular data point. It is a more elegant method of point location avoiding much of this time consuming process by reducing the number of reference points that must be spent to create data structures. Such structures range from particular orderings of reference points to trees in which reference points are organized into categories. The k-means algorithm uses a tree method to cluster three-dimensional data.

4.2 Cluster Error Determination Factor

As k-means approach is iterative, it is computationally intensive and hence applied only to images subareas rather than to full scenes and can be treated as unsupervised training areas. Consider a single cluster of points along with its centroid or mean. If the data points are tightly clustered around the centroid, the centroid will be representative of all the points in that cluster. The standard measure of the spread of a group of points about its mean is the variance or sum of the squares of the distance between each point and the mean. If the data points are close to the mean, the variance will be small. A generalization of the variance, in which the centroid is replaced by a reference point that may or may not be a centroid, is used in cluster analysis to indicate the overall quality of partitioning. Specifically, the error determination factor EDF is the sum of all the variances.

$$\text{EDF} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \| x_{ij} - z_i \|^2$$

Where $x_{ij}$ is the jth point in the ith cluster, $z_i$ is the reference point of the ith cluster and $n_i$ is the number of points in that cluster. The notation $\| x_{ij} - z_i \|$ stands for the distance between $x_{ij}$ and $z_i$. Hence, the error determination factor EDF indicates the overall spread of data points about their reference points. The error determination factor provides an objective method for comparing partitioning’s as well as a test for eliminating unsuitable partitioning’s. At present, finding the best partitioning requires generating all possible combinations of clusters and comparing their error measures. When clustering is done for the purpose of data reduction, as in the case of the remote sensing image dataset, the goal is to find the best partitioning. We merely want a reasonable consolidation of N data points into k clusters and if necessary, in this efficient way to improve the quality of the initial partitioning. All the data points are partitioned into k clusters by assigning each point to the cluster of the closest reference point. Adjustments are made by calculating the centroid for each of those clusters and then using those centroids as reference points for the next partitioning of all the data points. It can be proved that a local minimum of the error measure EDF corresponds to the reference point of its cluster than to any other reference point.

5 PROPOSED K-MEANS CLUSTERING ALGORITHM FOR REMOTE SENSING IMAGES

The basic aim is to classify colors in an automated fashion using the color space and k-means clustering. The entire process can be summarized in the following steps.

Step 1: Read the remote sensing image

Read the image from the remote sensing source is in jpeg format, which are Deforestation and SAR image ESL capabilities have expanded to include real-time reception and processing of data from LSU Earth Scan laboratory.

Step 2: For color separation of remote sensing image apply the decorrelation stretching method.

The primary purpose of decorrelation stretch is visual enhancement. The decorrelation stretching is a way to enhance the color differences in an image.

Step 3: Convert image from RGB color space to L*a*b* color space.

The L*a*b* color space (also known as CIELAB or CIE L*a*b*) enables you to quantify these visual differences. The L*a*b* space consists of a luminosity L* or brightness layer, chromaticity layer L*a* indicating where color falls along the red-green axis, and chromaticity layer L*b* indicating where the color falls along the blue-
yellow axis. All of the color information is in the \(a*\) and \(b*\) layers. We can measure the difference between two colors using the Euclidean distance metric. Covert the image to \(L^*a^*b^*\) color space.

Step 4: Classify the colors in a particular image of \(a^*b^*\) color space using K-Means Clustering method. Clustering is a way to separate groups of objects. The color based K-means clustering treats each object as having a location in space. It requires specifying the number of clusters to be partitioned and distances metric to quantify how close two objects are to each other. Since the color information exists in the \(a^*b^*\) space. Use k-means to cluster objects into five clusters using the Euclidean distance metric. The Euclidean distance is used as a metric and variance is used as a measure of cluster scatter.

Step 5: Label every pixel of the image using the results from K-means algorithm. For every object in the above input, K-Means returns an index corresponding to a cluster. Label every pixel in the image with its cluster index.

Step 6: Create images that segment the image by color. In computer vision, image segmentation is the process of partitioning a digital image into multiple segments (sets of pixels, also known as super pixels). The goal of segmentation is to simplify and/or change the representation of an image into something that is more meaningful and easier to analyze. Image segmentation is typically used to locate objects and boundaries (lines, curves, etc.) in images. More precisely, image segmentation is the process of assigning a label to every pixel in an image such that pixels with the same label share certain visual characteristics.

The result of image segmentation is a set of segments that collectively cover the entire image, or a set of contours extracted from the image (see edge detection). Each of the pixels in a region is similar with respect to some characteristic or computed property, such as color, intensity, or texture. Using pixel label, we have to separate objects in image by color, which will produce the result in five images.

Step 7: Arrange the nuclei of the image into a separate image window. The method uses examples selected by a user for building a statistical model that captures the texture and shape variations of the nuclei structures from a given image dataset to be segmented. Then programmatically determine the index of the cluster containing the blue object because k-means will not return the same cluster index value every time. We can do this using the cluster center value, which contains the mean \(a^*\) and \(b^*\) value of each cluster.

6 EXPERIMENTAL RESULTS

The various experiment carried out in the above remote sense imagery data set in MATLAB 7.6. The complete process and the standard are summarized in subsequent figure 2.
7 CONCLUSION

In general clustering algorithm is a well-known algorithm for classifying the Landsat images. But k-means clustering algorithm is also a well-known clustering algorithm to classify the remote sensing images. This algorithm splits the given image into different clusters of pixels in the feature space, each of them defined by its center. Initially each pixel in the image is allocated to the nearest cluster. Then the new centers are computed with the new clusters. These steps are repeated until convergence. We could assume random objects as the initial centroids or the first k objects in sequence could also serve as the initial centroids. These algorithms are robust and very effective in producing desired classifications especially in the field of pattern recognition as per the region of interest as demonstrated by the experimental results. This work presents a novel image segmentation based on colour features from the images. In this we did not used any training data and the work is divided into two stages. First enhancing color separation of satellite image using decorrelation stretching is carried out and then the regions are grouped into a set of five classes using K-means clustering algorithm. Using this two-step process, it is possible to reduce the computational cost avoiding feature calculation for every pixel in the image.

In future different neural network algorithms and fuzzy set algorithms can be used to classify the satellite images instead of above clustering algorithm and the classification results of those will be compared with the existing algorithms.

ACKNOWLEDGMENTS

We take this opportunity with much pleasure to thank my guide Dr. M Sumathi, Sri Meenakshi College for Women, Madurai -2, Tamilnadu and my colleagues for supporting this work.

REFERENCES


