Sindh Univ. Res. Jour. (Sci. Ser.) Vol.48 (2) 315-318 (2016)

SI NDH UNIVERSITYRESEARCH JOURNAL (SCIENCESERIES)

K-Means and ISODATA Clustering Algorithms for Landcover Classification Using Remote Sensing

A. W. ABBAS, N. MINALLH, N. AHMAD, S.A.R. ABID, M.A.A. KHAN

University of Engineering and Technology Peshawar, Pakistan

Received 17th June 2015 and Revised 28rd April 2016

Abstract-The aim of thisexploration work is to analyze the presentation ofunsupervised classification algorithms ISODATA(Iterative Self-Organizing Data Analysis Technique Algorithm)andK-Means in remote sensing, to evaluate statistically by iterative techniques to automatically group pixels of similar spectral features into unique clusters. This investigation used SUPARCO(Space and Upper Atmosphere Research Commission (Pakistan)) obtained remotely sensed patch of Abbottabad Pakistan. The test patch of Abbottabad is divided into Five bands i.e. NDVI (Normalized Difference Vegetation Index), green, near infrared, far infrared, and green. The ROIs (regions of interest) selected for classification of Land Cover data comprises five different types of classes i.e. water bodies, agriculture, settled area, forest and barren land. In this research of remote sensing the first step was to preprocess Abbottabad test patch by filtering, to improve performance of classification andneighboring pixels homogeneity. The next step was to assess the accuracy of Two pixel based unsupervised classifiers i.e. ISODATA and k-means on the said test patch. Finally, the mentioned classifiers performance is evaluated by varying their different parameters to categorize the effect of the clustering algorithms and their class statisticson whole classification outcomes.

Keywords: K-Means; ISODATA; Clustering Algorithms

1. INTRODUCTION

Remote Sensing (RS) imaginary is a vital source of information for observation of earth surface (RS). In modern terms, RS is the use of aerial sensor technologies to detect and classify objects from distance on Earth, its surface, atmosphere and oceans by means of propagated signals (Schowengerdt and Robert, 2007). RS can be divided into supervised and unsupervised classification, but mostly it is cumbersome to obtain prior knowledge due to the effect of image noise, various characteristics and complex background. Therefore unsupervised classification and cluster analysis is of great importance in RS studies (Wang and Cheng, 2010).The significance of Earth observation in our future decision-making processes through Remote Sensing, Pattern recognition, automatic classification, clustering, Change detection, feature extraction and parameter estimation is advantageous for economic reasons, disaster management, high yield of crop production, deforestation, security and surveillance (Schowengerdt and Robert, 2007), (Liu *et al.*, 2015).

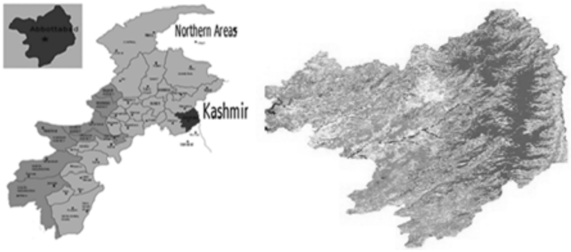
With advancement of technologies various classification approaches are deployed on remotely sensed images to get desired information, in (Venkateswaran*et al.*, 2013)Venkateswaran presented performance analysis of K-Means clustering for remotely sensed images. In this paper, a novel method for unsupervised classification in multi-temporal optical image based on DWT feature extraction and K-Means clustering is proposed. After preprocessing the optical image is feature extracted using the discrete wavelet transform. On the feature extracted image feature reduction is performed using energy based selection. Finally different K means clustering is performed and analyzed using MATLAB and ground truth data for improving classification accuracy. TaneeKamkhet in (Kamkh, 2012) discussed analysis of thaichote band characteristics using unsupervised pixel-based classification. Each band was individually classified using ISODATA and K-Means methods.In (Memarsadeghi*et al.*, 2003) an effective and modified version of ISODATA classifier is presented, which improved running time of classifier through storing of points in a KD-tree and estimates the dispersion of each cluster. Anil K. Jain in (Jain, 2010) praises K-Means clustering algorithm published in 1955 as the standard, simplest although many algorithms presented since then but K-Means is still widely used after 50 years. Therefore to design a general purpose clustering algorithm is cumbersome and difficult. The general issues in design of clustering algorithms, their overview, summary of clustering methods, and guidelines for latest research i.e. data clustering on large scale, real-time feature selection ensemble and semi-supervised are discussed.

The study area chosen for research isAbbottabad region of Hazara district in Khyber Pakhtunkhwa province of northeastern Pakistan with an altitudeof 1,260 meters (4,134 ft) and the total area of 1,967 square kilometers(SMEDA). The reasons for selection of this region are based on the area having 5types of land use with support for the interpretation

++Correspondence Author: Email Nasru Minallah <n.minallah@uetpeshawar.edu.pk>

these are agriculture, forest, settled area, barren and water (Raza*et al.*, 2012).The geographical position of Abbottabad region and its land patterns are shown in **"Fig.1 (a) and (b)".**

This paper illustrates the performance analysis of K-Means and ISODATA Clustering Algorithms for Landcover Classification. The continuing paper is structured into three sections. Section 2 illustrates Methodology. Performance analysis of results and discussions are elaborated in section 3while section 4 has presented conclusion and future work.



(a)(b)

FOREST AGRICULTUR WATER BARREN SETTELED

Fig1. (a)Abbottabad region geography(b)Land use pattern of Abbottabad

2. METHODOLOGY

Pattern recognition approaches are commonly deployed to recognize the underlying patterns in remotely sensed data (Websource). In this research the analysis started from data acquisition, pre-processing, then unsupervised classification by K-mean and ISODATA method was carried out, with final processing in post-classification and accuracy assessment has been discussed in next section using ENVI 5.0.

Data Acquisition

Datasets for RS and clustering can be from a wide range of sources like satellite sensor data, ground based sensor data, general data of weather, energy systems and so on. In this work dataset is the test patch of Abbottabad region KPK Pakistan acquired from satellite of SUPARCO. Once the dataset is acquired it is preprocessed, so that it is suitable for subsequent sub-processes.

Pre-processing

In this step using software ENVI 5.0, to obtain true color image load test patch in RGB color with sequence of green, red and near infrared bands. The test patch has5bandsi.e. red, green, far infrared, near infrared and one added band Normalized Difference Vegetation Index (NDVI). The true color image is pre-processed by filtering from convolutions and morphology using median filter. The filtered image is divided into 5types of lands i.e. agriculture, forest, settled area, barren land and water bodies using ROI tool.

Implementation of Unsupervised Classification By K-Mean And ISODATA Method

Currently various clustering algorithms are generally deployed in remote sensing. The two well-known are the K-Means and the ISODATA unsupervised classification algorithms.

These algorithms are iterative in nature. Firstly select the arbitrary starting values which show properties of cluster and effect result of classification.

Generally in both approaches first step is assignment of arbitrary initial values to cluster. Secondly classify each pixel to the nearby cluster. To calculate cluster mean of all pixels in one cluster is the third step. The repetition of 2nd and 3rd steps continues until the "change" between the iteration is small. The "change" can be considered in 2 ways either by the percentage of change of pixels from one iteration to another or by calculating the change of distances for the mean cluster vector between iterations.

In addition, for improvement the ISODATA consists of splitting and merging of clusters.

- The criteria for merging the clusters are based on certain threshold if the distance between the centers of 2 clusters is less than that or if the number of pixels in one cluster isfewer than that limit, clusters would be merged.
- The condition for splitting of clusters into 2 is satisfied if the cluster standard deviation increased than a predefined value and the number of pixels is 2 times the threshold for the minimum number of pixels(Tou and Gonzalez, 2012).

The K-Means Algorithm

Feature extraction is the most important step of any recognition system. The purpose of feature extraction is to take the important characteristics of the image and classify the overall image using this small set of information. The selection of features directly effects the classification operation. Good features results in a higher success rate in the process of recognition and vice versa. In this paper, two types of features have been extracted.

The goal of k-means is to reduce the variability within the cluster. The summation of squares distances termed as errors, between each pixel and its assigned cluster center is minimized and declared as objective function

$$\mathrm{SS}_{\mathrm{distances}} = \sum\nolimits_{\forall \mathrm{x}} [\mathrm{x} - \mathrm{C(x)}]^2 \; (1)$$

Wherein C(x) pixel x is assigned is to the mean of the cluster.

Mean Squared Error (MSE) is a measure of the within cluster variability and represented as.

$$\mathrm{MSE} = \frac{\sum_{\forall \mathrm{x}} [\mathrm{x} - \mathrm{C(x)}]^2}{(\mathrm{N} - \mathrm{c})^{\mathrm{b}}} = \frac{\mathrm{SS}_{\mathrm{distances}}}{(\mathrm{N} - \mathrm{c})^{\mathrm{b}}} (2)$$

In equation 2,b represents the number of spectral bands, whereas N is the number of pixels and c indicates the number of clusters(Websource1). K-Means implementation steps in ENVI 5.0 are shown in **(Fig.2).**

The ISODATA Algorithm

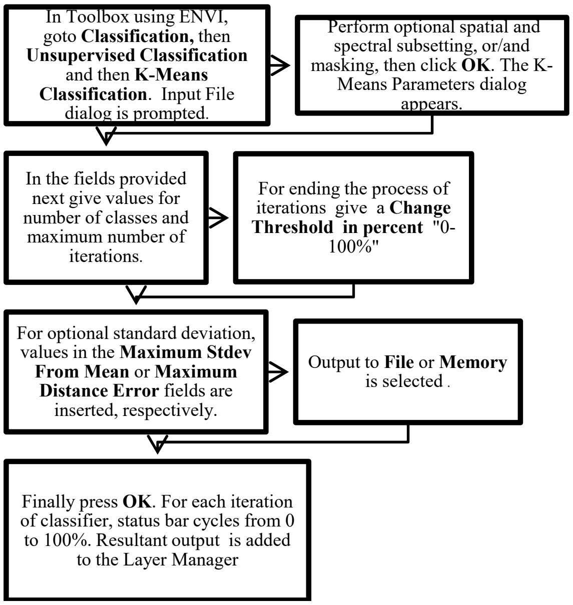
ISODATA computes class means consistently circulated in the data space before iteratively clusters the continuing pixels utilizing least distance approaches. Every iteration recalculates means as well as reclassifies pixels through respect to the new means, while in the K-Means approach, the number of clusters K remains the same throughout the iteration, although it mayturn out later that more or fewer clusters would fit the data better. This drawback can be overcome in the ISODATA Algorithm, which allows the number of clusters to be adjusted automatically during the iteration by merging similar clusters and splitting clusters with large standard deviations (Websource2). Implementation steps of ISODATA in ENVI 5.0 are shown in **(Fig.3)**

3. PERFORMANCE ANALYSIS OF RESULTS AND DISCUSSIONS

As shown in Fig. 1 (b) Abbottabad region is divided into 5types of land patterns but discussed in (Kamkhet. 2012) the overall of the accuracy by both the unsupervised classifiers is not the indicator for each land pattern. Therefore for cluster analysis, true color filtered image of Abbottabad has been used as test patch for the K-Means and ISODATA unsupervised classifiers by varying its parameters for performance analysis.

Using software Envi 5.0,**“**Fig. 4 (11)” shows true color median filtered image of Abbottabad region for further processing in the unsupervised data classification, “Fig. 4 (1 2)” is processed for K-Means with default values of number of classes 5 and change threshold of 5.0 as a result image is clustered into 5classes i.e. red, green, blue, cyan and yellow, while “Fig. 4 (1 3)” is clustering with ISODATA, with default initial values to insert in ENVI are classes = 5-10, iterations = 1, change threshold = 5.0, minimum pixels for class = 1, maximum class stdv= 1.000, class distance minimum = 5.000 and maximum merge pairs = 2, as a result 7classes of red, green blue, cyan, yellow, purple and indigo are formed.

“Fig. 4 (2 1)” and “Fig. 4 (2 2)” are obtained as a result of changing iterations from 1 to 10 for K-Means and ISODATA respectively. It has been seen that as iterations increases their accuracy increases, clusters and classes got uniform and clear and in ISODATA the resulted classes increased to maximum of 10.



In Toolbox using ENVI, goto **Classification,** then **Unsupervised Classification** and then **K-Means Classification**. Input File dialog is prompted.

Perform optional spatial and spectral subsetting, or/and masking, then click **OK**. The K-Means Parameters dialog appears.

In the fields provided next give values for number of classes and maximum number of iterations.

For ending the process of iterations give a **Change Threshold in percent** "0-100%"

For optional standard deviation, values in the **Maximum Stdev From Mean** or **Maximum Distance Error** fields are inserted, respectively.

Output to **File** or **Memory** is selected .

Finally press **OK**. For each iteration of classifier, status bar cycles from 0 to 100%. Resultant output is added to the Layer Manager

Fig 2. K-Means implementation work flow

By changing the number of classes from 5 to 10 for K-Means and 5-15 for ISODATA and keeping other parameters default, resulted images are “Fig. 4 (2 3)” and “Fig. 4 (3 1)” with 10classes respectively. It is deduced that when number of classes are same for both unsupervised algorithms then resulted images clusters are also same. Since it is justified that ISODATA is extension of K-Means and when number of classes’ parameter is same, their classification is also same.

Change threshold parameter is analyzed in “Fig. 4 (3 2)” and “Fig. 4 (3 3)” by varying it from 5% to 10% for K-Means and ISODATA respectively. It has been seen that there is no remarkable difference from default parameter and varied parameter in both the unsupervised algorithms.

After the classification processing, the result in each band was re-checked with the accuracy, by the existing land cover data, by cross classification and tabulation. In this, the yields should be classified into two parts: overall accuracy of the whole image and accuracy of each land cover.

In post classification, for class statistics of K-Means and ISODATA clustering with default parameter values has been used. Table 1 and Table 2 Class Means for each band number(class) against their values for full scene of 760,258points dimension has been evaluated for both the clustering algorithms. It has been shown that in both the cases class 1(red) has maximum class means comparatively.

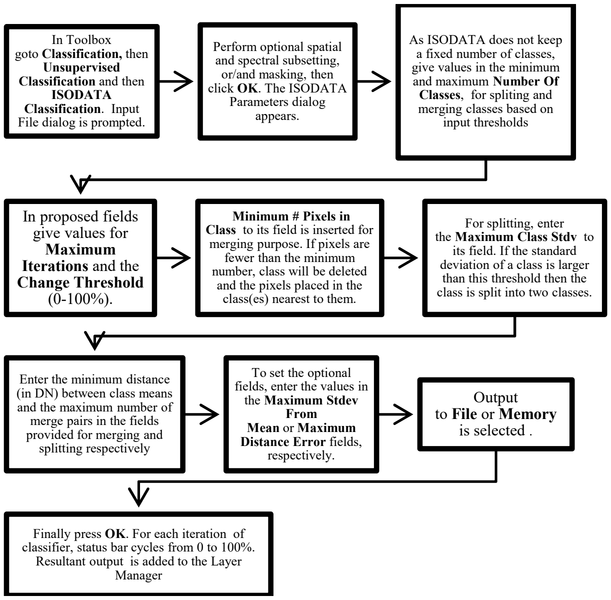

In Toolbox goto Classification, then Unsupervised Classification and then ISODATA Classification. Input File dialog is prompted.
Perform optional spatial and spectral subsetting, or/and masking, then click OK. The ISODATA Parameters dialog appears.
As ISODATA does not keep a fixed number of classes, give values in the minimum and maximum Number Of Classes, for spliting and merging classes based on input thresholds
In proposed fields give values for Maximum Iterations and the Change Threshold (0-100%).
Minimum # Pixels in Class to its field is inserted for merging purpose. If pixels are fewer than the minimum number, class will be deleted and the pixels placed in the class(es) nearest to them.
For splitting, enter the Maximum Class Stdv to its field. If the standard deviation of a class is larger than this threshold then the class is split into two classes.
Enter the minimum distance (in DN) between class means and the maximum number of merge pairs in the fields provided for merging and splitting respectively
To set the optional fields, enter the values in the Maximum Stdev From Mean or Maximum Distance Error fields, respectively.
Output to File or Memory is selected .
Finally press OK. For each iteration of classifier, status bar cycles from 0 to 100%. Resultant output is added to the Layer Manager

Fig 3. ISODATA implementation work flow

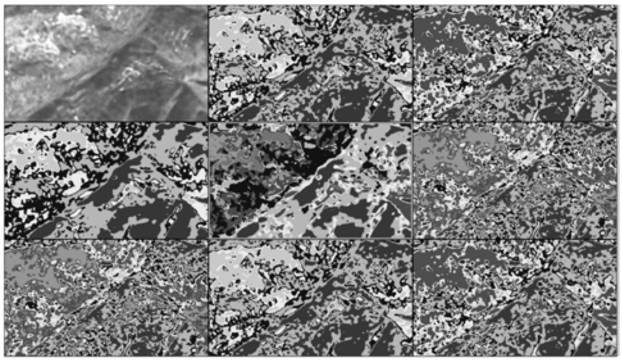


Fig 4. (1 1) Abbottabad region filtered image,

Table 1. Class Statistics for K-Means

Class/Band	Class Distribution	
	Occurrence of Individual Class (in Points)	*Occurrence of Individual Class (%age)*
Unclassified	0	00.000
Class 1 (Red)	197,431	25.699
Class 2 (Green)	177,599	23.117
Class 3 (Blue)	130,249	16.954
Class 4 (Yellow)	98,040	12.761
Class 5 (Cyan)	164,939	21.469

Table 2. Class Statistics for ISODATA

Class/Band	Class Distribution	
	Occurrence of Individual Class (in Points)	*Occurrence of Individual Class (%age)*
Unclassified	**0**	**00.000**
Class 1 (Red)	**160,351**	**20.872**
Class 2 (Green)	**134,212**	**17.470**
Class 3 (Blue)	**104,971**	**13.664**
Class 4 (Yellow)	**86,166**	**11.216**
Class 5 (Cyan)	**72,337**	**09.416**
Class 5 (Purple)	**59,037**	**07.685**
Class 5 (Indigo)	**151,184**	**19.679**

4. CONCLUSION

In a nutshell it has been concluded that unsupervised classification of land pattern with K-Means and ISODATA by varying the parameter values for number of iterations, number of classes and change threshold yields a different degree of accuracy and each justify its performance according to literature. In addition for post classification of class statistics, class 1(red) has maximum average.

In future, the performance of unsupervised algorithms with supervised algorithms will be analyzed to justify literature and their application according to situation.

REFERENCES:

Jain., A. K. (2010) "Data clustering: 50 years beyond K-Means," Journal of Pattern Recognition, Vol. 31, 651-666, Elsevier Science Inc. New York, NY, US.

Kamkhet.T, (2012)"Analysis of thaichote band characteristics using unsupervised pixel-based classification," 33rdAsian Conference on Remote Sensing (ACRS), Pattaya,Thailand, Vol. 1. 3:395- 401.

Liu.S., L. Bruzzone. F. Bovolo, M. Zanettiand P. Du. (2015) "Sequential Spectral Change Vector Analysis for Iteratively Discovering and Detecting Multiple Changes in Hyperspectral Images," Transactions on Geoscience and Remote Sensing, IEEE, Vol. 53, 8:4363–4378.

Memarsadeghi. N., N. Goddard. S. Nathan. and J. L, Moigne. (2003)"A fast implementation of the isodata clustering algorithm," IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Toulouse, France, Vol. 3, 2057–2059.

Raza. A., I. A. Raja. and S. Raza (2012) "Land-use change analysis of district abbottabad pakistan: taking advantage of gis and remote sensing."A scientific jour. of COMSATS – science vision Vol.18 No.1-2.675Pp.

Schowengerdt D.andA.Robert.(2007)"Remote sensing: models and methods for image processing,". Academic Press (3rd ed.) ISBN 978-0-12-369407-2, 02Pp.

Tou.J. T, and R. C.Gonzalez.(1994)."Pattern Recognition Principles,"Addison-Wesley Publishing Company, Reading, Massachusetts.

Venkateswaran. K., N. Kasthuri. K. Balakrishnan. and K. Prakash. (2013) "Performance Analysis of K-Means Clustering For Remotely Sensed Images," International Jour.of Computer Applications (0975–8887) Vol.84, 12.