JOURNAL OF ENVIRONMENTAL DESIGN (JED)

A Journal of the Faculty of Environmental Studies, University of Uyo, Uyo, Nigeria
Vol. 6, No. 2, Nov. 2011

TABLE OF CONTENT

EDITORIAL STYLE GUIDE FOR AUTHORS

Please submit your manuscript double spaced in MS Word 2003 - 2007 version. Provide tables and figures in a separate file (or files) in MS Excel MS Word, or EPS format. Maps should be supplied in EPS format.

Because manuscripts will undergo a blind review, submit two title pages. the first showing the title of the manuscript. author name, title. affiliation, telephone number. e-mail address. and the date of the manuscript. The second title page should contain only the title of the paper.

Third-person style is always preferred. If appropriate, authors may make limited use of first-person singular, but a single author should not refer to himself or herself as "we."

Biography: The manuscript should include, on a separate page or the "first" title page described above. a sentence listing each author's name, and affiliation.

Abstract. Include a one-paragraph abstract *not exceeding* 150 *words* and place it on the first page of the text. The abstract, describe the issue(s) or question(s) the paper addresses and state the major findings, conclusions and recommendations.

Keywords: To help users reference the JED published research, keywords are included with journal articles. Please suggest two keywords for your manuscript.

Abbreviations: The definition of an abbreviation or acronym is given the first time it appears; afterward, only the abbreviation is used. However, an abbreviation that is defined in the abstract should also be defined in the article. An abbreviation that appears only once in an article should be deleted and the full wording used.

If an abbreviation is first defined in the text, the abbreviation alone can then be used in subsequent footnotes or tables; however, if the abbreviation is first defined in a footnote or table, the abbreviation should be defined again when it first appears in the following text.

Text Headings: Headings are not numbered and are placed to the left. First-level headings are bold; second-level headings are italic; and third-level headings are italic with a period that leads directly into text.
Example:

> **First-** levelheading
> *Second-levelheading*
> *Third-levelheading.* The text continues ...

Tables and Figures: Use arabic numerals to number tables and figures consecutively in separate series in order of appearance. Include a brief descriptive title at the top of each. Tables and figures should be in separate page not integrated into the text. The text must contain a reference to each table not integrated into the text. The text must contain a reference to each table or figure.

Equations: Make sure that all symbols in equations are clear and that all equations (except those in footnotes) are numbered. Single-letter variables should be italicised. Multiple-letter variables and abbreviations (e.g., AGE) and functions (e.g., expo min. In) should not be italicised: neither should numbers. Parentheses, or mathematical operations. Vectors and matrices should be in bold (not italicized).

References: The manuscript must include complete and accurate citations of all materials referenced in the manuscript that are not of your original authorship. Please double-check your references to ensure that names and date are accurate, that Web pages are still active, and that there are no discrepancies between the text and the reference list. The APA style is strongly recommended.

UNSUPERVISED K-MEANS ALGORITHM FOR MULTI-SPECTRAL SATELLITE REMOTE SENSING IMAGE CLASSIFICATION

Onuwa Okwuashi[1] Mfon Isong[1], Aniekan Eyoh[1], Etim Eyo[2], and Aniekan D. Ekpo[3]
onuwaokwuashi@yahoo.com,

[1] Department of Geoinformatics & Surveying, University of Uyo. [2]School of Civil Engineering & Geosciences, Newcastle University, United Kingdom. [3]Amana Consortium Engineers Limited, Uyo.

ABSTRACT
This research explores the application of an unsupervised k-means algorithm to multi-spectral satellite remote sensing image classification, using a multi-spectral Landsat 7 ETM imagery of Porirua, New Zealand. MATLAB is used for implementing the k-means based computer program; while geographic information systems are used for data preparation and visualisation. The satellite image consists of three land use classes (water, undeveloped, and developed). No training of the k-means algorithm is required since k-means is an unsupervised classifier. The k cluster centroid locations and sums of point-to-centroid distances are first computed, and thereafter the distances from each point to every centroid. The classification solution for each pixel is found by determining the land use class that yields the least computed distance from each point to every centroid; such that the successful land use class wins the classification for that pixel. A total of 62,500 pixels are classified. The result of the experiment shows that not all the pixels are correctly classified. The classification result is validated with the Kappa statistic, based on a confusion matrix that compares the predicted with the referenced data. The calculated Kappa statistic is 0.8676, which indicates an almost perfect agreement between the predicted and the reference data.

INTRODUCTION

The process of relating pixels in remote sensing images to known land cover is called "image classification." The algorithms used to effect the classification process are called "image classifiers" (Mather, 1987). The extraction of land cover information from remote sensing images using image classifiers has been the subject of intense interest and research in the remote sensing community (Foody & Mather, 2004). Some of the traditional hard classifiers such as minimum distance to means and the box classifiers have been in use in remote sensing studies (Peddle, Foody, Zhang, Franklin, & LeDrew, 1994; Rogan, Franklin, & Roberts, 2002; Li, Chen, & Su, 2003; Mahesh & Mather, 2003). Because of the strong desire to maximise the degree of land cover information extracted from remotely sensed data research into new methods of classification has continued (Foody & Mather, 2004). The application of k-means algorithm to satellite remote sensing image classification problems is uncommon. The k-means algorithm is an unsupervised classification algorithm. Unsupervised classification means that no training examples are required to teach/train the classifier on how to classify a given data; instead the classifier uses cluster similarity to determine the most probable class of every pixel to be classified (Lo & Yeung, 2007). The objective of this research therefore is to illustrate how the k-means classifier can be applied to solving multi-class problems in satellite remote sensing image classification.

K-MEANS ALGORITHM

K-means (MacQueen, 1967) is one of the simplest unsupervised learning algorithms. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed *a priori*. The main idea is to define k centroids, one for each cluster. These centroids should be placed in a cunning way because different locations cause different results. Therefore, 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 an early groupage is done. At this point we need to re-calculate k new centroids as barycenters 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. In other words centroids do not move any more. Finally, this algorithm aims at minimizing an *objective function*, in this case a squared error function. The objective function is,

$$J = \sum_{j=1}^{k}\sum_{i=1}^{n}\left\| x_i^{(j)} - c_j \right\|^2 , \quad (1)$$

Where: $\left\|x_i^{(j)} - c_j\right\|^2$ is a chosen distance measure between a data point $x_i^{(j)}$; and the cluster centre c_j is an indicator of the distance of the n data points from their respective cluster centres (MacQueen, 1967).

The algorithm is composed of the following steps:

i. Place k points into the space represented by the objects that are being clustered. These points represent initial group centroids;

ii. Assign each object to the group that has the closest centroid;

iii. When all objects have been assigned, recalculate the positions of the K centroids;

iv. Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimised can be calculated.

Although it can be proven that the procedure will always terminate, the k-means algorithm does not necessarily find the most optimal configuration, corresponding to the global objective function minimum. The algorithm is also significantly sensitive to the initial randomly selected cluster centres. The k-means algorithm can be run multiple times to reduce this effect (MacQueen, 1967).

APPLICATION

A multi-spectral Landsat 7 ETM imagery of Porirua, New Zealand, acquired in 2006 was used for the experiment (see Figure 1). The Landsat image consists of seven spectral bands, and has a cell size of 25m x 25m. The original satellite data were first reviewed in GIS (ArcGIS software), and all seven spectral bands were extracted using the *layer properties tool* and visualised in MATLAB (see Figure 1). Before importing the data into MATLAB, they were first converted from raster to ASCII data using the ArcGIS *conversion tool*. MATLAB cannot read raster files; hence the data must be in ASCII format for onward processing in MATLAB. In MATLAB the final study area was extracted from the original satellite image. Some regions of the satellite image are affected by cloud, which was why the final study area did not include the regions affected by cloud. The final image used for the classification was 250 250 pixels, which amounts to 62,500 pixels. All the seven spectral bands were used for the classification. The satellite image consists of three distinct land use classes: water, undeveloped, and developed cells. The aim of this experiment therefore is to classify the satellite image into these three land use classes.

Since k-means is an unsupervised algorithm, the algorithm was not trained on how to classify the data unlike in the case of supervised algorithms. First, using the MATLAB function *[IDX,C] = kmeans(X,k) the k* cluster centroid locations in the k-by-p matrix C were computed (see Table 1); second, using the MATLAB function *[IDX,C,sumd] = kmeans(X,k)* the within-cluster sums of point-to-centroid distances in the 1-by-k vector sumd were computed (see Table 2); and third, using the MATLAB function *[IDX,C,sumd,D]. = kmeans(X,k)* distances from each point to every centroid in the n-by-k matrix D were computed (see Table 3). Where *IDX* represents the classification indices for the three land use classes *(water=1; undeveloped=2; developed=3)*; *C* represents the computed *k* cluster centroid locations; *X* represents the input data (that is, all the data from the extracted 7 bands); *k* represents the three land use classes (water, undeveloped, and developed), therefore the numerical value of k was 3; *sumd* represents the computed sums of point-to-centroid distances; and *D* represents computed distances from each point to every centroid. The classification results were visualised in ArcGIS (see Figure 2).

From Table 3, the computed distances from each point to every centroid determine the final classification result. A total of 62,500 pixels were classified. All the 62,500 results cannot be displayed, hence 28 results were displayed. Some of the pixels were wrongly classified when the k-means results were compared with the reference data.

CONCLUSION

The result of the classification experiment displayed in Figure 2 was validated with the Kappa statistic (Cohen, 1960).). Kappa statistic can be expressed mathematically as:

$$k = \frac{P_o - P_c}{1 - P_c} \quad (2)$$

Where,

$$P_o = \sum_{i=1}^{m} P_{ii} = \frac{1}{N}\sum_{i=1}^{m} n_{ii} \quad (3)$$

and,

$$P_c = \sum_{i=1}^{m} P_{i+}P_{+i} = \frac{1}{N^2}\sum_{1=1}^{m} n_{i+}n_{+i} \quad (4)$$

(Ma & Redmond, 1995; Lo & Yeung, 2007).

Where,

P_o = proportion agreement observed

P_c = proportion agreement expected by chance

n_{ii} = the total number of correctly classified points by class along the diagonal of the error matrix

N = the total number of points checked (sampled)

P_{ii} = the proportion of correctly classified sample points by class at the diagonal of the error matrix (i.e. n_{ii} / N)

P_{i+} = the marginal distribution of the sample data (n_{i+} / N where n_{i+} is the row sum by class)

P_{+i} = the marginal distribution of the reference data (n_{+i} / N where n_{+i} is the column sum of class)

m = the total number of classes

The Kappa statistic is more reliable than other validation techniques because it has the ability to evaluate the actual agreement and chance agreement (Fung & LeDrew, 1988). Kappa statistic is computed from an *error matrix* or *confusion matrix* resulting from the comparison of the reference with the predicted data (see Table 4). A cell-by-cell comparison between the reference and the predicted data are displayed in the confusion matrix given in Table 4. The computed Kappa statistic using equations 2, 3, and 4 was 0.8676.

CONCLUSION

Unsupervised classifiers such as the k-means employ simple and less cumbersome algorithms in resolving classification problems. The painstaking selection of training samples (like in supervised classification) is avoided. The calculated Kappa statistic indicates that the predicted data are almost in perfect agreement with the reference data. Even though supervised classifiers are preferred to unsupervised classifiers, the result of this experiment has showed that unsupervised classifiers can equally furnish reliable results when applied to satellite remote sensing image classification problems.

REFERENCES

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37-46.

Foody, M. G. & Mather, A. (2004). Toward intelligent training of supervised image classifications: directing training data acquisition for SVM classification. *Remote Sensing of Environment, 93*, 107-117.

Fung, T. & LeDrew, E. (1988). The Determination of optimal threshold levels for changes detection using various accuracy indices. *Photogrammetric Engineering and Remote Sensing, 54*(10), 1449-1454.

Landis, J. R. & Koch, G. G (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*,159 – 174

Li, T. S., Chen, C. Y., & Su, C. T. (2003). Comparison of neural and statistical algorithms for supervised classification of multi-dimensional data. *Int. J. Indus. Eng.—Theory Appl. Pract., 10*, 73–81.

Lo, C. P. & Yeung, A. K. W. (2007). *Concepts And Techniques Of Geographic Information Systems* (2nd ed.). Upper Saddle River, NJ: Pearson Prentice Hall.

Ma, Z. & Redmond, R. L. (1995). Tau coefficients for accuracy assessment of classification of remote sensing data. *Photogrammetric Engineering and Remote Sensing 61*, 435-439.

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, University of California Press, *1*, 281-297.

Mahesh P. & Mather, P. M. (2003). An assessment of the effectiveness of decision tree methods for land cover classification. *Remote Sens. Environ., 86*, 554–565.

Mather, P. (1987). *Computer Processing of Remotely Sensed Images – An Introduction*. New York: John Wiley and sons.

Peddle D. R., Foody, G. M., Zhang, A., Franklin, S. E., & LeDrew, E. F. (1994). Multisource image classification II: An empirical comparison of evidential reasoning and neural network approaches. *Canadian Journal of Remote Sensing, 20*, 396–407.

Rogan, J., Franklin, J., & Roberts, D. A. (2002). A comparison of methods for monitoring multitemporal vegetation change using Thematic Mapper imagery. *Remote Sens. Environ., 80*, 143–156.

Table 1: *Computed K cluster centroid locations*

	Band 1	Band 2	Band 3	Band 4	Band 5	Band 6	Band 7
Water	74.2838	26.96785	27.34722	82.37702	67.44065	116.5622	21.81936
Undeveloped	80.68658	27.53548	21.94871	13.17946	9.702583	109.8495	2.208192
Developed	87.98545	36.08719	40.76915	100.0753	106.5402	121.1811	40.77096

Table 2: *Computed sums of point-to-centroid distances*

Land use class	Sums
Water	24181447
Undeveloped	2406576
Developed	48820532

Table 3: *Some computed distances from each point to every centroid for 62,500 pixels (Water=1; Undeveloped=2; Developed=3)*

Pixel	Water	Undeveloped	Developed	K-means Classification	Reference data	Remark
1	21567.41	192.2854	10126.24	1	1	Correct
2	21937.53	209.6763	10410.35	1	1	Correct
3	22164.16	239.5595	10602.28	1	1	Correct
4	21937.53	209.6763	10410.35	1	1	Correct
5	21695.94	184.7469	10246.37	1	1	Correct
---	---	---	---	---	---	---
7774	20357.63	143.3535	9322.284	3	2	Wrong
7775	20842.18	264.1524	9496.304	3	3	Correct
7776	15166.17	449.3151	6070.433	3	2	Wrong
---	---	---	---	---	---	---
16710	2034.605	2[illegible]19.81	5891.427	2	2	Correct
16711	996.9755	20253.32	4517.526	2	3	Wrong
16712	1069.494	17962.12	3550.998	3	3	Correct
16713	884.6756	18787.14	3557.691	3	2	Wrong
16714	514.8906	16722.61	1630.35	3	3	Correct
16715	1254.799	14679.31	812.1255	3	2	Wrong
16716	578.3176	16558.65	1349.07	1	1	Correct
16717	337.19	17985.51	1902.765	1	2	Wrong
16718	708.888	13189.97	691.0484	3	3	Correct
16719	1765.036	9222.489	558.5814	3	3	Correct
16720	1342.327	10860.98	1235.95	3	3	Correct
---	---	---	---	---	---	---
62492	9414.154	2420.056	2305.833	2	2	Correct
62493	3103.303	9445.394	191.3742	2	2	Correct
62494	904.9282	21058.65	2859.661	2	2	Correct
62495	868.0736	22489.45	3528.721	2	2	Correct
62496	1033.212	19669.93	2347.416	2	2	Correct
62497	1214.678	18311.07	1850.91	2	2	Correct
62498	1276.014	15332.9	1001.424	2	2	Correct
62499	1133.715	16721.4	1354.674	2	2	Correct
62500	585.832	21132.13	2865.911	2	2	Correct

Table 4: *Computed confusion matrix for k-means classification*

	REFERENCE DATA		
	Developed	Undeveloped	Water

PREDICTED DATA			
Developed	15897	2061	0
Undeveloped	1229	35251	0
Water	161	1084	6817

According to Landis and Koch (1977) the computed Kappa result can be appraised based on the interpretation given in Table 5. The computed Kappa statistic implies an almost perfect agreement with the reference data.

Table 5: *Interpretation of kappa statistic*

KAPPA	INTERPRETATION
< 0	No agreement
0.0 - 0.20	Slight agreement
0.21 – 0.40	Fair agreement
0.41 – 0.60	Moderate agreement
0.61 – 0.80	Substantial agreement
0.81 – 1.00	Almost perfect agreement

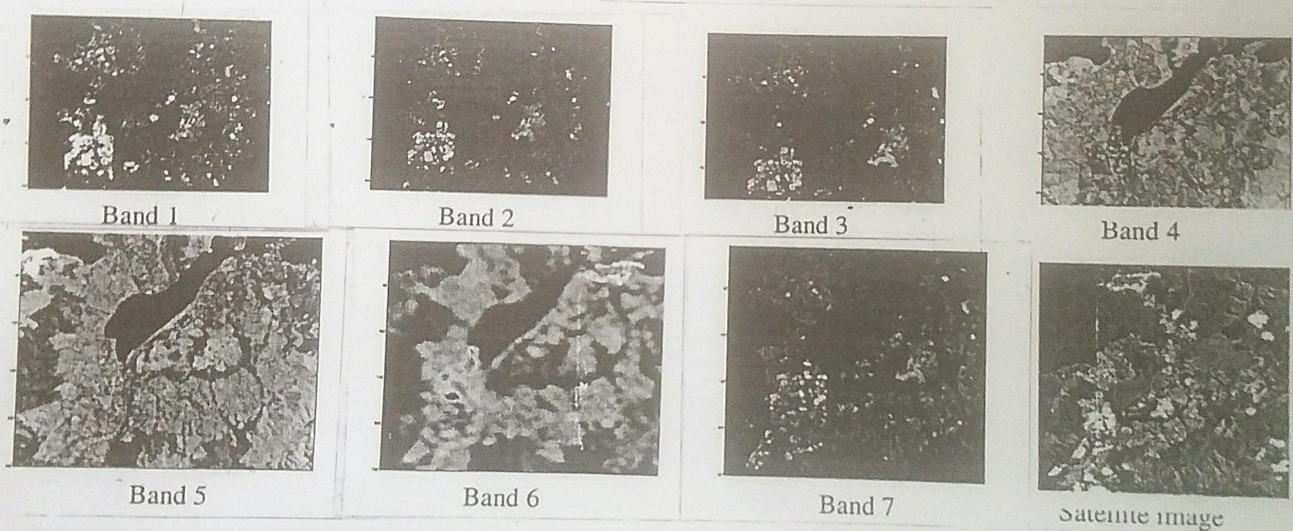

Figure 1: *Extracted bands 1 - 7 of Landsat image of Porirua and original Landsat image of Porirua, New Zealand*

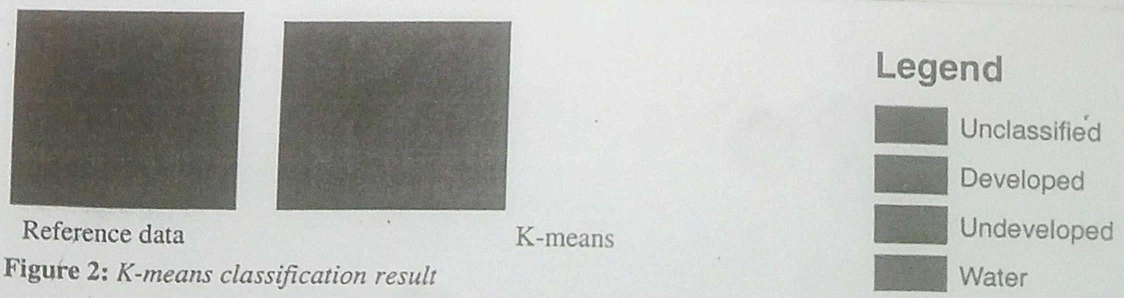

Figure 2: *K-means classification result*