ISSN No. 0976-5697

Volume 4, No. 10, Sep-Oct 2013

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Canberra Distance based Approach for Classifying Remote Sensing Images using Affinity Propagation

D.Napoleon
Assistant Professor
Department of Computer Science
Bharathiar University
mekaranapoleon@yahoo.co.in

M. Praneesh
Assistant Professor
Department of Computer Science
Sankara College of Science and Commerce
raja.praneesh@gmail.com

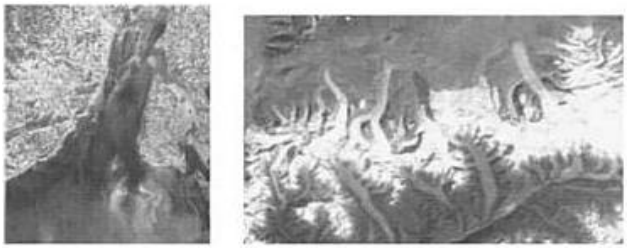
Abstract - Clustering is the process of subdividing an input data set into a desired number of subgroups so that members of the same subgroup are similar and members of different subgroups have diverse properties. Many heuristic algorithms have been applied to the clustering problem, which is known to be NP Hard. This paper represents a frame work for clustering on image data. The objective of the frame work algorithm used to perform clustering on very large data sets, especially on Remote sensing image data sets. Efficient time techniques are used as a performance measure for clustering on image data.

Key Words: Data Mining, Clustering, K-means algorithm, Genetic algorithm, Image data

I. INTRODUCTION

Data mining techniques are the result of a long process of research and product development [2]. This evolution began when business data was first stored on computers, continued with improvements in data access, and more recently, generated technologies that allow users to navigate through their data in real time. Data mining takes this evolutionary process beyond respective data access and navigation to prospective and proactive information delivery. Data mining is ready for application in the business community because it is supported by three technologies such as Massive data collection, Powerful multiprocessor computers and Data mining algorithms [11]. Data mining deals with large databases that impose on clustering analysis additional severe computational requirements. Cluster analysis is mainly conducted on computers to deal with very large scale and complex datasets. With the development of computer based techniques, clustering has been widely used in data mining, ranging from web mining, image processing, machine learning, artificial intelligence, pattern recognition, social network analysis, bioinformatics, geography, geology, biology, psychology, sociology, customers behavior analysis, marketing to e-business and other fields[13].

Information regarding the natural resources, such as agricultural, hydrological, mineral, forest, geological resources, etc., can be extracted based on remotely sensed image analysis. For remotely sensed scene analysis, images of the earth's surface arc captured by sensors in remote sensing satellites or by a multi-Spectra) scanner housed in an aircraft and then transmitted to the Earth Station for further processing [3, 4]. We show examples of two remotely sensed images in Figure 1 whose color version has been presented in the color figure pages. Figure 1(a) shows the delta of river Ganges in India. The light blue segment represents the sediments in the delta region of the river, the deep blue segment represents the water body, and the deep red regions are mangrove swamps of the adjacent islands. Figure 1.1(b) is the glacier flow in Bhutan Himalayas. The white region shows the stagnated ice with lower basal velocity.



(a) (b)

Figure. 1: Example of a remotely sensed image of (a) delta of river Ganges, (b) Glacier flow in Bhutan Himalayas

The task of grouping data points into clusters of "similar" items are a form of unsupervised learning that has applications in many fields. For instance, current techniques used for machine vision require processing of digital information obtained from pixels [2]. A very important step in this digital information processing is to group the data in some fashion so that patterns can be recognized. Clustering can be used for this task.

II. RELATED WORK

Clustering has become a widely studied problem in a variety of application domains, such as in data mining and knowledge discovery [1], [2] statistical data analysis [3], [4] data classification and compression [6], medical image processing [5] and bioinformatics [6]. Several algorithms have been proposed in the literature for clustering [7], [8]. A. L. Abul is explained about Cluster Validity Analysis Using Sub sampling [10]. The objective of all clustering algorithms is to divide a set of data points into subsets so that the objects within a subset are similar to each other and objects that are in different subsets have diverse qualities [11], [12], [13]. Bradley and Fayyad have proposed an algorithm for refining the initial cluster centers. Not only are the true clusters found more often, but the clustering algorithm also iterates fewer times. Some clustering

methods improve performance by reducing the distance calculations. For example, Judd *et al.* proposed a parallel clustering algorithm P-CLUSTER which uses three pruning techniques. Ming-Chuan Hung, explained about an Efficient *k*-Means Clustering Algorithm Using Simple Partitioning.

The Genetic algorithm described in uses a multi step procedure. The authors refer to this procedure as a semi supervised form of learning. In [19] a GA is used to solve the clustering problem for a data set of geographical data. Similarly, Yan-He Chen describes about Genetic algorithm for Aerial image clustering.

III. FRAME WORK

The proposed architecture is designed for the classification of remote sensing images which is based on land cover information. The proposed work has been systematically represented the following figure

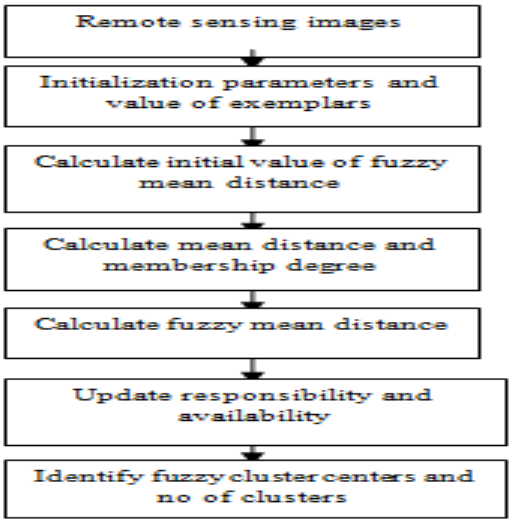


Figure-2 System Architecture

The Canberra distance is a metric function often used for data scattered around an origin. It was introduced in 1966 (Lance & Williams 1966) and is today mainly used in the form of 1967 (Lance & Williams 1967). The Canberra metric is similar to the Manhattan distance (which itself is a special form of the Minkowski distance). The distinction is that the absolute difference between the variables of the two objects is divided by the sum of the absolute variable values prior to summing. The generalized equation is given in the form:

$$d^{CAD}(i,j) = \sum_{k=0}^{n-1} \frac{|y_{i,k} - y_{j,k}|}{|y_{i,k}| + |y_{j,k}|}$$

This is a slightly modified form compared to the original form given by Lance & Williams (1966) and was suggested by Adkins (reference in Lance & Williams 1967). In the equation d^{CAD} is the Canberra distance between the two objects *i* and *j*, *k* is the index of a variable and *n* is the total number of variables *y*. This approach seems to be more suitable for the distance based clustering. The Affinity Propagation Algorithm:

A. Initialization:

$r(i,k) = s(i,k) - \max \{ s(i,k') \}$, $a(i,k) = 0 \quad k' \neq k$

B. Responsibility updates:

$$r(i,k) = s(i,k) - \max_{k' s.t. k' \neq k} \{ a(i,k') + s(i,k') \}$$

$$r(i,k) = s(i,k) - \max_{k' s.t. k' \neq k} \{ a(k,k') + s(k,k') \}$$

C. Availability updates:

$a(i,k) = \min (0, r(k,k) + \sum_{i' s.t. i' e \{I,k\}} \max \{ 0, r(i',k) \})$

$a(k,k) = \sum_{i' s.t. i' e \{I,k\}} \max \{ 0, r(i',k) \})$

D. Making assignments

$C^*_i \leftarrow \arg\max_{1 \le k \le n} r(i,k) + a(i,k)$.

In this processing, two kinds of messages are exchanged among data points, and each takes into account a different kind of competition. **:** Let X = { x1,x2,……xn} $\subset R^p$ be a set of pixels vectors, where X represents all pixels in the data set, $x_i = [x_{i1}, x_{i2}, \ldots\ldots, x_{ip}]$ (i= 1,2,….,n) is the feature vector of pixel i, n is the total number of pixels in the image, and p is the number of features considered (e.g., bands of the image). Let Z = { x1,x2,……xm} $\subset R^p$ be a set of cluster exemplars, where Z represents all exemplars set, $z_k = [z_{k1}, z_{k2}, \ldots\ldots z_{kp}]$ (k= 1,2 …m) is an exemplar vector, and m is the number of clustering exemplars in the image (at initialization n = m). $N=\{ \mu_1(1), \mu_2(2), \ldots \mu_n(m)\}$ is the membership degree set, and $\mu_i(k) = [\mu_{i1}(k), \mu_{i2}(k) \ldots\ldots, \mu_{ip}(k)]$ (i= 1,2,…..n; k=1,2…..m) is a membership degree vector.

IV. EXPERIMENTAL RESULTS

The dataset used in our experiments is a portion of remote sensing image of Quick bird images, which covers a small area of the south part of the city of Trento, italy acquired on july 17, 2006. This site mainly contains two land-cover types, which are vegetation and exposed land. In this dataset, green areas represent vegetation, dark areas represent dry land, slightly darker green areas represents grass land, and deep darker areas are represent paddy field. We can see that the degree of class mixture in the vegetation area is high. A brown area mainly represents exposed land; slightly brown areas mainly represent dry salt flats which are blocked by forest land and dry land that are highly mixed too.

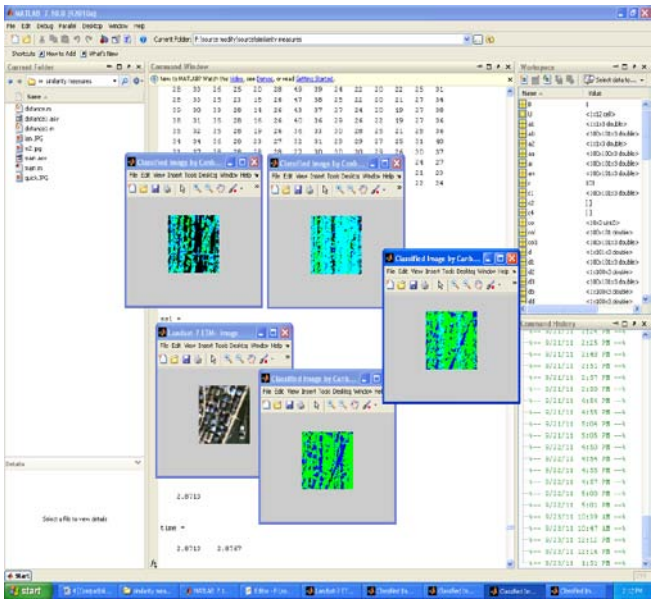


Figure-3 Results of Canberra Distance Approach

Table -1 Error Matrices

Class	Agricultural field	road	tree	soil	roof	shadow	Grass	Row total
Agricultural field	**136**	0	1	7	0	0	0	144
Road	0	**104**	0	0	16	5	0	125
Tree	0	0	**34**	1	18	2	0	55
Soil	3	0	3	**58**	5	0	0	69
Roof	0	0	3	2	**81**	2	2	90
Shadow	2	2	2	13	13	**99**	0	131
Grass	9	1	2	5	0	3	**35**	55
Column total	150	107	45	86	134	111	37	**547**
Producer accuracy	89.33	93.26	75.11	60.28	57.21	80.19	89.19	
User accuracy	92.48	86.12	57.26	73.00	80.66	75.79	63.46	

The clustering results of Canberra distance based approach are evaluated using overall accuracy, Kappa-value, average of Producers accuracy, average user accuracy, overall accuracy and kappa value are widely used in the validation 2. J. R. Wen, J. Y. Nie, and H. J. Zhang, "*Query clustering using user logs*," ACM Transactions on Information Systems, Vol. 20, 2002, pp. 59-81.

Of the land use and land cover Classification. The following table represents execution time and proposed frame work.

Table -2 Accuracy

Parameter	Proposed Approach
Execution members	23
Execution Time	52
Overall accuracy	92.12
Kappa Value	0.531

Kappa coefficients are widely used as Classification accuracy assessment for remote sensed data. The result of performing kappa analysis is a KHAT statistic (an estimate of Kappa), which is computed as

$$\hat{K} = \frac{N\sum_{i=1}^{r} x_{ii} - \sum_{i=1}^{r}(x_{i+}.x_{+i})}{N^2 - \sum_{i=1}^{r}(x_{i+}.x_{+i})}$$

Where r is the number of rows in the error matrix (also called the confusion matrix), x_{ii} is the number of observations in row i and column$i$, x_{i+} and x_{+i} are the marginal totals of row i and column$i$, respectively, and N is the total number of observations.

V. CONCLUSION

In this work a fast and efficient remote sensing classification system based on land cover information was proposed frame work to classify the images. This work illustrates and concluded that the system performance has very good efficiency and high accuracy than traditional clustering algorithm. In future we are improve our accuracy based on some statistical methodologies will be implemented. The classified images can be compared with ground truth information physically.

VI. REFERENCES

[1] Z. Huang, "Extensions to the k-means algorithm for clustering large data sets with categorical values," Data Mining and Knowledge Discovery, Vol. 2, 1998, pp. 283-304.

[2] J. Banfield and A. Raftery, "Model-based gaussian and non-gaussian clustering," Biometrics, Vol. 49, 1993, pp. 15-34.

[3] J. L. Bentley, "Multidimensional binary search trees used for associative searching," Communications of the ACM, Vol. 18, 1975, pp. 509-517.

[4] D.A. Clausi, "K-means iterative fisher unsupervised clustering algorithm applied to image texture segmentation," Pattern Recognition, Vol. 35, 2002, pp. 1959-1972.

[5] F. X. Wu, W. J. Zhang, and A. L. Kusalik, "Determination of the minimum samples size in micro array experiments to cluster genes using K-means clustering," in Proceedings of 3rd IEEE Symposium on Bioinformatics and Bioengineering, 2003, pp. 401-406.

[6] K. Alsabti, S. Ranka, and V. Singh, "An efficient k-means clustering algorithm," in Proceedings of 1st Workshop on High performance Data Mining, 1998.

[7] R. C. Dubes and A. K. Jain,"Algorithms for Clustering Data", Prentice Hall, 1988.

[8] E. R. Ruspini, "A new approach to clustering," Inform. Contr., vol. 19, pp. 22–32, 1969.

[9] L. Abul, R. Alhajj, F. Polat and K. Barker "Cluster Validity Analysis Using Sub sampling," in proceedings of IEEE International Conference on Systems, Man, and Cybernetics, Washington DC, Oct. 2003 Volume 2: pp. 1435-1440.

[10] J. Grabmeier and A. Rudolph, "Techniques of cluster algorithms in data mining," Data Mining and Knowledge Discover, 6, 2002, pp. 303-360.

[11] L. O Hall, I. B. Ozyurt, J. C. Bezdek, "Clustering with a genetically optimized approach," IEEE Transactions on Evolutionary Computation,3(2), 1999, pp. 103-112.