ISSN No. 0976-5697

Volume 3, No. 2, March-April 2012

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Optimized Clustering Algorithm for Content-Based Image Retrieval

Dr. A.V. Senthil Kumar*
Director, Department of Computer Applications,
Hindusthan College of Arts & Science, Coimbatore.
avsenthilkumar@yahoo.com

J.Sivakami, Research Scholar
Department of Computer Applications,
Hindusthan College of Arts & Science, Coimbatore.
sivakamimscss@gmail.com

Dr. Ibrahiem El Emary
King Abdul Aziz University,
Jeddah, Kingdom of Saudi Arabia
omary57@hotmail.com

Dr.Adnan Alrabea
Albalqa Applied University,
Jordan
adnan_alrabea@yahoo.com

Abstract: Content-Based Image Retrieval (CBIR), is mainly based on finding images of interest from a large image database using the visual content of the images. Most of the approaches to image retrieval were text-based, where individual images had to be annotated with format. Existing works are based on the performance of a number of clustering algorithms in image retrieval has been analyzed. The proposed work in this paper is viewed on a new fuzzy based c-means partitional clustering algorithm. Partitional clustering algorithm is used to improve the Content Based Image Retrieval and for comparing the performance of the image content.

Keywords: Clustering, Partitional algorithm, CBIR, Fuzzy C-means.

I. INTRODUCTION

Knowledge discovery in databases process, or KDD is relatively young and interdisciplinary field of computer science is the process of discovering new patterns from large datasets involving methods at the intersection of artificial intelligence, machine learning, statistics and database system [1]. The goal of data mining is to extract knowledge from a data set in a human-understandable structure. Data mining is the entire process of applying computer-based methodology, including new techniques for knowledge discovery [2], from data. Databases, Text Documents, Computer Simulations, and Social Networks are the sources of data for mining.

Clustering is a method of unsupervised classification, where data points are grouped into clusters based on their similarity. The goal of a clustering algorithm is to maximize the intra-cluster similarity and minimize the inter-cluster similarity [3]. Clustering is a data mining (machine learning) technique used to place data elements into related groups without advance knowledge of the group definitions [3]. A cluster is a collection of data points that are similar to one another within the same cluster and dissimilar to data points in other clusters [4].

Partitional and hierarchical clustering are the most widely used forms of clustering. In partition clustering, the set of n data points are partitioned into k non-empty clusters, where $k \leq n$. In the case of hierarchical clustering, the data points are organized into a hierarchical structure [5], resulting in a binary tree or dendogram. In this paper, we propose a new clustering algorithm, which would come under the category of partitional clustering algorithms. Two commonly used methods for partitioning data points include the k-means method and the k-medoids method. In the k-means method, each cluster is represented by its centroid or the mean of all data points in the cluster [6]. In the case of the k-medoids method, each cluster is represented by a data point close to the centroid of the cluster. Apart from these methods, there has been lots of work on fuzzy partitioning methods and partition methods for large scale datasets [7].

II. REVIEW OF LITERATURE

Cluster analysis or clustering is the task of assigning a set of objects into groups (called clusters) so that the objects in the same cluster are more similar (in some sense or another) to each other than to those in other clusters [4], [8]. Clustering is a main task of explorative data mining, and a common technique for statistical data analysis used in many fields, including machine learning, pattern recognition, image analysis [8],informationretrieval, andbioinformatics. Popular clustering techniques include k-means clustering and Expectation Maximization (EM) clustering.

A clustering is essentially a set of such clusters, usually containing all objects in the data set [8]. Additionally, it may specify the relationship of the clusters to each other, for example a hierarchy of clusters embedded in each other. Clustering can be roughly distinguished in:

a. Hard clustering: each object belongs to a cluster or not
b. Soft clustering (also fuzzy clustering): each object belongs to each cluster to a certain degree (e.g. a likelihood of belonging to the cluster)

We use the notion of 'contribution of a data point' for partitional clustering. The resultant algorithm requires only three passes and we show that the time complexity of each pass is same as that of a single iteration of the k-means clustering algorithm. While the k-means algorithm optimizes only on the intra-cluster similarity, our algorithm also

optimizes on the inter-cluster similarity. Clustering has widespread applications in image processing. Color-based clustering techniques have proved useful in image segmentation [9]. The k-means algorithm is quite popular for this purpose. Clustering based on visual content of images is an area that has been extensively is in research for several years. This finds application in image retrieval.

Content-Based Image Retrieval (CBIR) aims at finding images of interest from a large image database using the visual content of the images. Traditional approaches to image retrieval were text-based, where individual images had to be annotated with textual descriptions. Since this is a tedious manual task, image retrieval based on visual content is very essential [10].

III. EXISTING SYSTEM

Clustering is a form of unsupervised classification that aims at grouping data points based on similarity. In this paper, they propose a new partitional clustering algorithm based on the notion of 'contribution of a data point'. They apply the algorithm to content-based image retrieval and compare its performance with that of the k-means clustering algorithm [2].

Partitional clustering aims at partitioning a group of data points into disjoint clusters optimizing a specific criterion. When the number of data points is large, a brute force enumeration of all possible combinations would be computationally expensive. Instead, heuristic methods are applied to find the optimal partitioning [10]. The most popular criterion function used for partitional clustering is the sum of squared error function given by a widely used squared-error based algorithm is the k-means clustering algorithm. In this paper, we propose a clustering algorithm similar to the k-means algorithm. They define the contribution of a data point belonging to a cluster as the impact that it has on the quality of the cluster. This metric is then used to obtain an optimal set of '*k*' cluster from the given set of data points.

They now outline the proposed contribution-based clustering algorithm. It optimizes on two measures, namely the intracluster dispersion given by

$$\alpha = \frac{1}{n} \sum_{x \in C_i} (x - m_i)^2$$

And the inter-cluster dispersion given by

$$\beta = \frac{1}{k} \sum_{i=1}^{k} (m_i - \bar{m})^2$$

Where *k* is the number of clusters and is the mean of all centroids. The algorithm tries to minimize α and maximize β.

IV. THE FUZZY C-MEANS ALGORITHM

In this paper, a new fuzzy based C- means partitional clustering algorithm based on the notion of 'contribution of a fuzzy data point'. Is proposed the algorithm to Content Based Image Retrieval and compare its performance with that of the partitional clustering algorithm is applied. During this, the clustering accuracy will be improved and the number of iteration to form clustering is also reduced. By applying 'contribution of a fuzzy data point', the time of evaluation is minimized.

The Fuzzy C-Means algorithm is the most popular objective function based fuzzy clustering algorithm. The FCM was first developed by [8].The objective function used in FCM is given by equation

$$J_{FCM}^{m}(U, A, X) = \sum_{i=1}^{c} \sum_{j=1}^{n} \mu_{ij}^{m} d_{ij}^{2} = \sum_{i=1}^{c} \sum_{j=1}^{n} \mu_{ij}^{m} \left\| \underline{x}_j - \underline{a}_i \right\|^2$$

Where $\mu_{ij} \in [0,1]$ is the membership degree of data object $\underline{x}_j$ in cluster C_i, and it satisfies the following constraint given by equation

$$\sum_{i=1}^{c} \mu_{ij} = 1, \forall j = 1, 2, \ldots, n$$

C is the number of clusters, m is the fuzzifier, m >1, which controls the fuzziness of the algorithm. These two parameters need be specified before running the algorithm. The measure 2 $d_{ij}^2 = \left\| \underline{x}_j - \underline{a}_i \right\|^2$ is the square Euclidean distance between data object $\underline{x}_j$ to center $\underline{a}_i$.

Minimizing the objective function with the constraint is a non-trivial constraint nonlinear optimization problem with continuous parameters $\underline{a}_i$ and discrete parameters μ_{ij}. So there is no obvious analytical solution. Therefore, an alternative optimization scheme, which optimizes one set of parameters while the other set of parameters are considered as fixed, is used here. Then the updating functions for $\underline{a}_i$ and μ_{ij} is obtained as shown in equation. The algorithm is described in the following.

Step 1: Determining the number of cluster, c, m-value (let m = 2), and the converging error, $\varepsilon > 0$ (such as $\varepsilon = 0.001$) and choosing the initial membership matrix:

$$U^{(0)} = \begin{bmatrix} \mu_{11}^{(0)} & \mu_{12}^{(0)} & \ldots & \mu_{1n}^{(0)} \\ \mu_{21}^{(0)} & \mu_{22}^{(0)} & \ldots & \mu_{2n}^{(0)} \\ \ldots & \ldots & & \ldots \\ \mu_{c1}^{(0)} & \mu_{c2}^{(0)} & \ldots & \mu_{cn}^{(0)} \end{bmatrix}$$

Step 2: To calculate

$$\underline{a}_i^{(k)} = \frac{\sum_{j=1}^{n} \left[\mu_{ij}^{(k-1)} \right]^m \underline{x}_j}{\sum_{j=1}^{n} \left[\mu_{ij}^{(k-1)} \right]^m} \quad i = 1, 2, \ldots, c$$

$$\mu_{ij}^{(k)} = \left[\sum_{l=1}^{c} \left[\frac{\left(\underline{x}_j - \underline{a}_i^{(k)}\right)' \left(\underline{x}_j - \underline{a}_i^{(k)}\right)}{\left(\underline{x}_j - \underline{a}_l^{(k)}\right)' \left(\underline{x}_j - \underline{a}_l^{(k)}\right)} \right]^{\frac{1}{m-1}} \right]^{-1}$$

Step3: Increment k until $\max_{1 \le i \le c} \left\| \underline{a}_i^{(k)} - \underline{a}_i^{(k-1)} \right\| < \varepsilon$

V. EXPERIMENTAL RESULTS

The images were clustered using our algorithm with the initial centroids chosen at random. The cluster whose centroid was closest in distance to the given test image was determined and the images belonging to the cluster were retrieved. The results were then compared with images retrieved using the kmeans clustering algorithm with the same setofinitialcentroids.

In this Figure 1, the basic partitional clustering algorithm for Content-Based Image Retrieval is compared with the proposed fuzzy algorithm partitional clustering for Content-based image retrieval. When the number of clusters increases, the average precision will be increased. The average precision

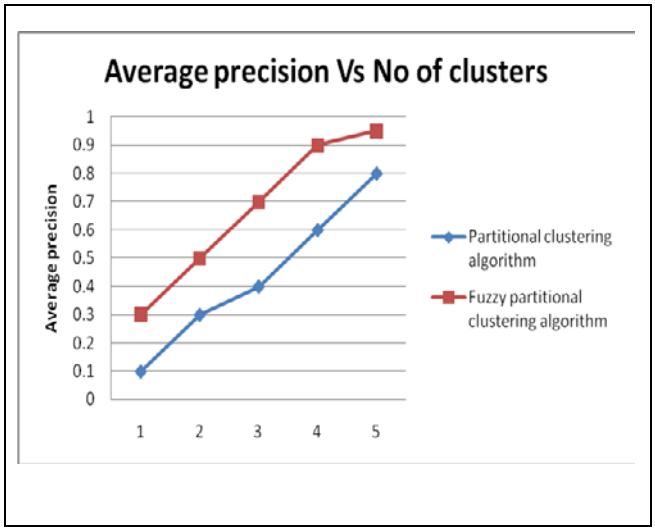

Figure1 :Average precision and Number of a Clusters

Rate varies in the interval of 0.1. Fuzzy partitional clustering algorithm gives better results than the existing clustering algorithm. The number of clusters increased by 1.

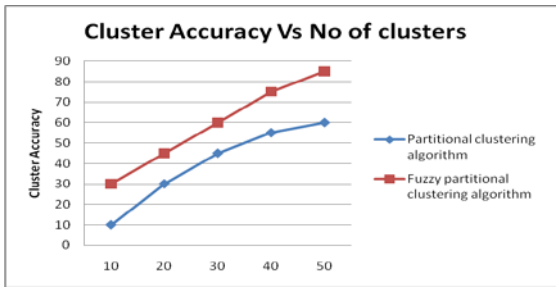

Figure2:Cluster accuracy and number of clusters

In this Figure 2, the basic partitional clustering algorithm for Content-Based Image Retrieval is compared with the proposed fuzzy algorithm partitional clustering for Content-based image retrieval. When the number of clusters increases, the Cluster accuracy will increased. Cluster accuracy increases by 10s than the existing system. Here the number of clusters increased by 10. Thus the cluster accuracy is improved in the proposed system than the existing system.

VI. CONCLUSION

In this paper, we have thus proposed a new Fuzzy C-Means partitional clustering algorithm based on the notion of 'contribution of a data point' .Unlike the k-means algorithm, our algorithm optimized on both the intra-cluster and inter-cluster similarity measures and required fewer passes with each pass having the same time complexity as that of the k-means algorithm. Organizing the retrieved search results into clusters is an intuitive form of content representation and facilitates user's browsing of images. While the performance of a number of clustering algorithms in image retrieval has been analyzed in existing paper, we apply our proposed algorithm to CBIR and compare its performance with that of the k-means clustering algorithm. We applied the clustering algorithm to content-based image retrieval and our experiments reveal that the algorithm improves on recall at the cost of precision.

VII. REFERENCES

[1]. R. Xu and D. Wunsch, "Survey of clustering algorithms," IEEE Transactions on Neural Networks, Vol.16, Issue 3, pp. 645– 678, May 2005.

[2]. P. Bradley, U. Fayyad, and C. Reina, "Scaling clustering algorithms to large databases," in Proc. 4th Int. Conf. Knowledge Discovery and Data Mining (KDD'98), 1998, pp. 9–15.

[3]. F.H. Long, H.J. Zhang, and D.D. Feng, "Fundamentals of content-based image retrieval," in D.D. Feng, W.C. Siu, and H.J. Zhang (Eds), 'Multimedia information retrieval and management—technological fundamentals and applications', Springer-Verlag, New York, 2003, pp. 1–26

[4]. Y. Liu, D. Zhang, G. Lu, and W. Ma, "A survey of content-based image retrieval with high-level semantics," PatternRecogn., vol. 40(1), 2007, pp. 262-282.

[5]. D. Cai, X. He, Z. Li, W. Ma, and J. Wen, "Hierarchical clustering of WWW image search results using visual, textual and link information," in Proc. 12th Annual ACM International Conference on Multimedia (MULTIMEDIA '04), Oct. 2004, New York, NY, pp. 952-959.

[6]. H. Zhang, J.E. Fritts, and S.A. Goldman, "Image segmentation evaluation: A survey of unsupervised methods," Comput. Vis. Image Underst., 110(2), May. 2008, pp. 260-280.

[7]. P.J. Dutta, D.K. Bhattacharyya, J.K. Kalita and M. Dutta, "Clustering approach to content based image retrieval," in Proc. Conference on Geometric Modeling and Imaging: New

Trends (GMAI), 2006, IEEE Computer Society, Washington, DC, pp. 183-188

[8]. Y. Chen, J.Z. Wang, and R. Krovetz, "Content-based image retrieval by clustering," in Proc. 5th ACM SIGMM international Workshop on Multimedia information Retrieval, Berkeley, California, Nov. 2003, MIR '03, ACM, New York, NY, pp. 193-200.

[9]. K. Jarrah, S. Krishnan, L. Guan , "Automatic content-based image retrieval using hierarchical clustering algorithms," in Proc. International Joint Conference on Neural Networks (IJCNN '06), Oct. 2006, Vancouver, BC pp. 3532 - 3537.

[10]. "Object and concept recognition for content-based image retrieval," [Online]. Available: http://www.cs.washington.edu/research/imagedatabase/groundtruth/