Advances in Computing 2016, 6(1): 1-5
DOI: 10.5923/j.ac.20160601.01

Fuzzy K-means Application to Semantic Clustering for Image Retrieval

Enikuomehin A. O.[1,*], Rahman M. A.[1], Zubair A. F.[2], Ahmed A.[2], Ogundipe A. O.[3], Amin A. O.[4], Egbudin D. M.[4]

[1]Department of Computer Science, Lagos State University, Lagos, Nigeria
[2]Department of Computer Science, Nothingham Trent University, Nottingham, UK
[3]Department of Computer Science, Crawford University, IGBESA, Nigeria
[4]Department of Computer Science, University of Ilorin, Ilorin, Nigeria

Abstract Several approaches have been used in defining the semantic features of images. This paper considers the most efficient approach by comparing the high points of kmean algorithm and fuzzy k mean algorithms. We observe that both approaches are efficient however based on the experimental set up, result shows that hidden features of images such as color texture and shape are more captured using fuzzy based k mean approach. The paper concludes by recommending larger experimental setup for further testing.

Keywords K-means, Semantic Clustering, Information Retrieval

1. Introduction

Clustering techniques have been used in several areas of research essentially when the data set seems to be enormous and not truly structured. Clustering models are increasingly used in finding appropriate method for solving complex problems when classification is required for data items acquired in an unsupervised manner. A general approach in clustering involves the classification of data set into groups. The data in the same subclass or group will have peculiar characteristic features that unite them and also distinguishes them from data items in other subclasses. Clustering has a wide area of application which cut across science, social sciences, management sciences, medical sciences amongst others [1], [2].

In this paper, the interest is in image processing vis as vis classification technique for effective retrieval. We basically consider the best methods for achieving image clustering using a K- and Fuzzy K- means approaches. Image clustering has made segmegation more effective and the resulting applications include the ability to segment a still or moving image including movie. The principal approach in clustering is the unsupervisory method by which user can select data points without a guide. The technique has widely been accepted as a time saving method thus, its huge acceptability. We present an approach suitable for the combined use of the K means for image clustering. We first highlight the features of each approach and then justify the need for a combination by providing a usable algorithm.

* Corresponding author:
toyinenikuomehin@gmail.com (Enikuomehin A. O.)
Published online at http://journal.sapub.org/ac

2. Image Feature

The data points of an image are used in the classification of the image. This is due to the semantic features of the images. The semantic analysis of images is attracting many researchers in the area of image retrieval [3] though many advances have been made especially in the area of color but little has been achieved in the other areas such as shape and texture [4], [5]. To effectively develop an image retrieval approach, methods for extracting some knowledge for such images must also be considered. As in image mining that is being used in computer vision, image processing, artificial intelligence amongst others, it's important to present algorithms that are being implemented by rules. This rule base system will have the capabilities of identifying low level and high level features of an image. This identification is suitable for points clustering.

3. K-Means

K-means was introduced by James MacQueen in 1967 [6]. It is observed that a lot of work has been done in this field. In the time frame of 1967 to 1998, all the research work was related to the introduction of K-means in clustering area. After this, all the modifications and improvements were started on K-means clustering. Newton and Mitra [7] used

fuzzy clustering along with a neural network thus making a hybrid architecture known as the Adaptive Fuzzy Leader Clustering (AFLC). In the control structure of the neural network the Fuzzy K-Means learning algorithm is embedded. The empirical results revealed that the hybrid architecture is capable of arbitrary data patterns. Hui Xiong, Junjie Wu and Jian Chen [8] studied K-Means in data – distribution perspective. They described many validation measures such as CV, purity, entropy and F-measure. They came to know that K-Means produces clusters of uniform size. The true cluster sizes might be slightly different even. Dehariya, Shailendra and Jain [9] presented experimental results of both K-Means and Fuzzy K-Means for image segmentation and proved that Fuzzy K-Means is better than K-Means due to the quality of the result provided by the Fuzzy K-Means approach. De-Sheng and Ming-Qin [10] applied Fuzzy K-Means algorithm for getting required information from the Internet. The search engine results can be clustered with satisfactory performance. The clustering takes place based on sentence similarity. Yang [11] has made an extensive survey of cluster analysis that involved both K-means and FKM algorithms. It does mean that it explored both soft and hard clustering. From the review, they concluded that the fuzzy clustering algorithms will have obvious performance gains over their counterparts that do not use fuzzy logic. However, they also consume more computational resources of the system.

Vuda Sreenivasarao [12] has developed a model for improving academic performance evaluation of students based on data warehousing and data mining techniques that use soft-computing intensively. Valdés-Pasarón, Márquez and Ocegueda-Hernández [13] proposes a methodology using fuzzy logic to measure the quality of education by using quantitative and qualitative values with the hopes to develop criteria for the quality of education in a way closer to the realities of Latin American countries. Oyelade, Oladipupo and Obagbuwa [14] have implemented K-means clustering algorithm to analyze the academic performances of students on the basis of some pre set measures and the Euclidean distance as a measure of similarity distance which provide a simple and qualitative methodology to compare the predictive power of clustering algorithm was adopted.

4. Brief on Fuzzy Logic and Fuzzy Set Theory

The real world is complex; this complexity generally arises from uncertainty. Humans have unconsciously been able to address complex, ambiguous, and uncertain problems, thanks to the gift of thinking. This thought process is possible because humans do not need the complete description of the problem since they have the capability to reason approximately. With the advent of computers and their increase in computation power, engineers and scientists are more and more interested in the creation of methods and techniques that will allow computers to reason with uncertainty and vagueness.

Fuzziness is a language concept; its main strength is its vagueness using symbols and defining them. Consider a set of tables in a lobby, in classical set theory, we would ask: Is it a table? And we would have only two answers, *yes* or *no*. If we code *yes* with a 1 and *no* with a 0, then we would have the pair of answers as (0, 1). At the end we would collect all the elements with 1 and have the set of tables in the lobby. We may then ask what objects in the lobby can *function* as a table? We could answer that tables, boxes, desks, among others can function as a table. The set is not uniquely defined, and it all depends on what we mean by the word *function*. Words like this have many shades of meaning and depend on the circumstances of the situation. Thus, we may say that the set of objects in the lobby that can't function as a table is a *fuzzy set*, because we have not crisply defined the criteria to define the *membership* of an element to the set. Objects such as tables, desks, boxes may function as a table with a certain degree, although the fuzziness is a feature of their representation in symbols and is normally a property of models, or languages.

Membership functions are mathematical tools for indicating flexible membership to a set, modeling and quantifying the meaning of symbols. They can represent a subjective notion of a vague class, such as chairs in a room, size of people, and performance among others. Commonly there are two ways to denote a fuzzy set. If X is the universe of discourse, and x is a particular element of X, then a fuzzy set A defined on X may be written as a collection of ordered pairs:

$$A = \{(x, \mu_A(x))\} \quad : \quad x \in X$$

where each pair $(x, \mu_A(x))$ is a singleton.

5. Choosing between K-Means and Fuzzy K-Means Clustering Technique for Solving Complex Interrelated Problems

Clustering is one of the data mining techniques that have been around to discover business intelligence by grouping objects into clusters using a similarity measure. Data Mining is the analysis of datasets that are observational, aiming at finding out unsuspected relationships among datasets and summarizing the data in such a noble fashion that are both understandable and useful to the data users [7]. Clustering is an unsupervised learning process that has many utilities in real time applications in the fields of marketing, biology, libraries, insurance, city-planning, earthquake studies and document clustering. Latent trends and relationships among data objects can be unearthed using clustering algorithms. Many clustering algorithms came into existence. However, the quality of clusters has to be given paramount importance.

K-Means has been around for many years to discover patterns by grouping objects based on some similarity measure. It is faster and simple. However, it takes uniform

clusters and needs to know the number of clusters beforehand. Another important feature of K-Means is that it keeps an object into a specific cluster. However, in the real world an object might be closer to more than one cluster. The K-Means clustering is also known as hard clustering. K-Means is an algorithm to classify or to group your objects based on attributes/features into K number of group. K-Means or Hard C-Means clustering is basically a partitioning method applied to analyze data and treats observations of the data as objects based on locations and distance between various input data points. Partitioning the objects into mutually exclusive clusters (K) is done by it in such a fashion that objects within each cluster remain as close as possible to each other but as far as possible from objects in other clusters. K is positive integer number. The grouping is done by minimizing the sum of squares of distances between data and the corresponding cluster centroid. Thus the purpose of K-mean clustering is to classify data. Each cluster is characterized by its centre point i.e. centroid. The distances used in clustering in most of the times do not actually represent the spatial distances. In general, the only solution to the problem of finding global minimum is exhaustive choice of starting points. In a dataset, a desired number of clusters K and a set of k initial starting points, the K-Means clustering algorithm finds the desired number of distinct clusters and their centroids. A centroid is the point whose co-ordinates are obtained by means of computing the average of each of the co-ordinates of the points of samples assigned to the clusters.

K-means clustering is most widely used clustering algorithm which is used in many areas such as information retrieval, computer vision and pattern recognition. K-means clustering assigns *n* data points into *k* clusters so that similar data points can be grouped together. It is an iterative method which assigns each point to the cluster whose centroid is the nearest. Then it again calculates the centroid of these groups by taking its average. K-means algorithm has the following steps precisely.

Simply put, k-Means Clustering is an algorithm among several that attempt to find groups in the data. The pseudo code follows the following procedure.

Initialize $\mathbf{m}_i$, i = 1,…,k, for example, to k random $\mathbf{x}^t$
Repeat
  For all $\mathbf{x}^t$ in X
    $b_i^t \leftarrow 1$ if $\| \mathbf{x}^t - \mathbf{m}_i \| = \min_j \| \mathbf{x}^t - \mathbf{m}_j \|$
    $b_i^t \leftarrow 0$ otherwise
  For all $\mathbf{m}_i$, i = 1,…,k
    $\mathbf{m}_i \leftarrow$ sum over t $(b_i^t \mathbf{x}^t)$ / sum over t (b_i^t)
Until $\mathbf{m}_i$ converge

The vector **m** contains a reference to the sample mean of each cluster. **x** refers to each of our examples, and **b** contains our "estimated [class] labels".

This can be explained as follow;

1. Form initial centroids based on the number of clusters (K)
2. Assign each object taken from the data set to the nearest centroid based on the data set distance from the centroid to complete the initial grouping process.
3. Then recalculate the new cluster centroids by the average of all data points that are assigned to the clusters.
4. Repeat the steps 2 and 3 until there is no need for the centroids to be adjusted. Thus, the final clusters are formed.

Fuzzy K-Means which is known as soft clustering approach came into existence to overcome the limitations of k-means. According to YE Ping [8], Fuzzy K-Means is an improved form of K-Means algorithm which allows the degree of belonging. This improvement makes Fuzzy k-Means flexible enough and allows an object to belong to more than one cluster. Fuzzy K-Means has better utility in the real world applications than K-Means with respect to the quality of clusters. Coefficients are used to provide the degree of belongingness and they are defined as follows;

$$x\sum_{k=1} num.clusters \ \mu(x)=1$$

In case of Fuzzy K-Means the mean of all points constitute the centroid. The objects are weighted by the degree in which they belong to a particular cluster.

$$\text{Center}_K = \sum_x \mu k(x) mx \ / \sum_x \mu k(x) m$$

Fuzzy K-Means has the following algorithm in execution precisely;

1. Choosing number of clusters
2. Assigning coefficients of points randomly for being in the clusters
3. Every time the coefficients' change between two iterations is observed and the sensitivity threshold is considered.
4. This process continues until the convergence of the algorithm.

6. Implementation

Three basic semantic features of images are used in the experimentation of the proposed model. The features are color, texture and shape respectively. To extract blue color, a process of segmentation must first be carried, using [15] whose idea presents that an easier method for image segmentation is to disambiguate the RGB image into PQR as a form of expanded chromaticity, where P reprecent Perceived Brightness, Q; blue –yellow and R represent red-green. Other conventional colors such as red, blue, yellow, purple, torkorish, etc are also identified and a linear combination of them. This will mean that there can be a linear transformation of any colour to one of the 180 reference colours either as one to one or as many to one. We can now use the K-means algorithm to perform clustering.

The outcome of the operation is the generation of X region where similar color is fused as clusters.

Similarly, to identify the texture characteristics of the image, we use the Quasi Gabor filter [16] which has so far proven reliable for some feature identification. [17] [18] The image is classified according to the energy level and graded in the conventional frequency grades of f=n*2 where starting n is 1 thereby having 2,4,8,16,32 etc as the values of f. The orientation in degrees is identified and an average scores for each block taken. The image s are transformed into values which represent a set of straight line, curves and arcs. This is achieved by transforming the image into binary using B_spline and Bezier curve. This approach has also been suitably used in polygonal approximation.

These features are then used in the formulation of an appropriate database for retrieval and experimental processes. Initiating the retrieval process, color properties such as contrast, light dark, simultaneous contrast e.t.c. were identified for each x and a production rule is defined. For texture, cluster centres were used to define the fuzzy rules and finally the shape used fuzzy production rule to calculate between the similarity of search shape and a given object.

7. Experimental Results

We hereby present the result of the experimental runs. Though many runs were used in the experiment but only 20 can be effectively reported due to space and set up accuracy. We ensure the randomness of the sample over the same range of 6 to 14 is maintained with varying sample size. The Standard deviation and mean are reported. Varied SD was allowed in the first set of runs while the second set was fixed to 2.

Set 1: Variable Standard Deviation										
Run:	1	2	3	4	5	6	7	8	9	10
k:	6	7	11	6	11	6	10	7	7	8
Iterative:	3	11	2	7	8	8	6	3	22	6
Recursive:	8	8	9	8	9	8	8	8	8	8

t 2: Fixed Standard Deviation = 2										
Run:	1	2	3	4	5	6	7	8	9	10
k:	9	7	5	11	11	9	7	7	6	10
Iterative:	8	8	14	21	10	11	7	4	2	5
Recursive:	8	8	8	9	8	8	8	8	8	8

Figure 1. Experimental result

The data shows the correctness of the algorithm in successful determination of K. Issues arises when the clusters are few and the iteration is short, the algorithm performs well considering that some of the means are incredible close.

8. Conclusions

Data classification has been an important area in many natural sciences field. Thus, the popularity of data clusters will not be a surprise to data scientist. Most algorithms, including the simple K-means, are admissible algorithms. This paper present algorithms suitable for image portioning based on their semantic features. Experimental set up was used to justify the performance of the algorithm in identifying a sample K cluster. Result shows that fuzzy k means performs better with limited data set as used in this paper. The outcome is impressing and the model show great prospect in the advancement of image retrieval and content analysis domain.

REFERENCES

[1] Nasaroui, Olfa, Fabio Gonzalez, and Dipankar Dasgupta. "The fuzzy artificial immune system: Motivations, basic concepts, and application to clustering and web profiling." In Fuzzy Systems, 2002. FUZZ-IEEE'02. Proceedings of the 2002 IEEE International Conference on, vol. 1, pp. 711-716. IEEE, 2002.

[2] Agrawal, Rakesh, Johannes Gehrke, Dimitrios Gunopulos, and Prabhakar Raghavan. Automatic subspace clustering of high dimensional data for data mining applications. Vol. 27, no. 2. ACM, 1998.)

[3] Campbell, Walter S., James R. Campbell, William W. West, James C. McClay, and Steven H. Hinrichs. "Semantic analysis of SNOMED CT for a post-coordinated database of histopathology findings." Journal of the American Medical Informatics Association 21, no. 5 (2014): 885-892.

[4] Lu, Zhiwu, Liwei Wang, and Ji-Rong Wen. "Direct Semantic Analysis for Social Image Classification." In Twenty-Eighth AAAI Conference on Artificial Intelligence. 2014.

[5] Blaschke, Thomas, Geoffrey J. Hay, Maggi Kelly, Stefan Lang, Peter Hofmann, Elisabeth Addink, Raul Queiroz Feitosa et al. "Geographic object-based image analysis–towards a new paradigm." ISPRS Journal of Photogrammetry and Remote Sensing 87 (2014): 180-191

[6] MacQueen, James. "Some methods for classification and analysis of multivariate observations." In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, vol. 1, no. 14, pp. 281-297. 1967.

[7] Newton, S.C., Mitra, S., (1992). An Adaptive Fuzzy System For Control And Clustering Of Arbitrary Data Patterns. IEEE. Page (383-370).

[8] Hui Xiong, Junjie Wu, Jian Chen. (2009). K-Means Clustering Versus Validation Measures., A Data-Distribution Perspective. IEEE. Vol.39., No. 2, Page (318-331).

[9] Vinod Kumar Dehariya, Shailendra Kumar Shrivastava and Jain R.C., (2010). Clustering of Image Data Set Using K-Means And Fuzzy K-Means Algorithms. IEEE. Page (386-391).

[10] Zhu De-Sheng and Zhou Ming-Qin., (n.d)., Get What You Want from Internet Using Fuzzy k-means Clustering Algorithm. IEEE. Page (1-4).

[11] Yang M.S., (1993). A Survey of Fuzzy Clustering, Mathl. Computational Modeling., Vol. 18., No. Page (1- 16).

[12] Vuda, S., & Yohannes, G. (2012). Improving Academic Performance of Students of Defense University Based on data Warehousing and Data Mining. Global Journal of Computer Science and Technology. Vol. 12., No. 2., Page (201-209).

[13] Valdés-Pasarón S., Márquez B.Y., and Ocegueda-Hernández J.M., Methodology for Measuring the Quality of Education Using Fuzzy Logic. Software Engineering and Computer Science, 2011, Volume 180, No. 5., Page (509-515).

[14] Oyelade O.J., Oladipupo O.O., Obagbuwa I.C., (2010) Application of k-Means Clustering algorithm for prediction of Students' Academic Performance. (IJCSIS) International Journal of Computer Science and Information Security, Vol. 7.

[15] Han, Jia-Wei, Jian Pei, and Xi-Feng Yan. "From sequential pattern mining to structured pattern mining: a pattern-growth approach." Journal of Computer Science and Technology 19, no. 3 (2004): 257-279.

[16] Park, Mira, Jesse S. Jin, and Laurence S. Wilson. "Fast content-based image retrieval using quasi-gabor filter and reduction of image feature dimension." InImage Analysis and Interpretation, 2002. Proceedings. Fifth IEEE Southwest Symposium on, pp. 178-182. IEEE, 2002.

[17] Turner, Mark R. "Texture discrimination by Gabor functions." Biological cybernetics 55, no. 2-3 (1986): 71-82.)

[18] Idrissa, Mahamadou, and Marc Acheroy. "Texture classification using Gabor filters." Pattern Recognition Letters 23, no. 9 (2002): 1095-1102.