Republic of Iraq

Ministry of Higher Education and Scientific Reseach
University of Babylon
College of Education for Pure Sciences

K-means clustering

**Research Submitted to university of Babylon / college of Education for Pure Sciences / Mathematic Department as part of The Requirements for
The Degree of B.Sc. in mathematical science**

By

Ruwaida Raed Thahir

Supervised by

Asst. Dr. Amera Abdul Wohid Funjan

1444/1443

2023/2022

إقرار المشرف

أشهد بأن إعداد هذا المشروع الموسوم

(~~~~~~~~~K-means clustering~~~~~~~~)

والمعد من قبل الطالبة (رويده رائد ظاهر)

قد تم تحت إشرافي في قسم الرياضيات / كلية التربية للعلوم الصرفة / جامعة بابل

التوقيع:

الاسم:

المرحلة العلمية:

التاريخ:

بَسِمً آلَلَہ آلَرحًمًنِ آلَرحًيَمً

(وَعْدَ اللَّـهِ لَا يُخْلِفُ اللَّـهُ وَعْدَهُ وَلَـكِنَّ أَكْثَرَ النَّاسِ لَا يَعْلَمُونَ*يَعْلَمُونَ ظَاهِرًا مِّنَ الْحَيَاةِ الدُّنْيَا وَهُمْ عَنِ الْآخِرَةِ هُمْ غَافِلُونَ).

"سورة الروم، آية: 6-7".

آلَآهّدٍآء

الى من نصح و ارشد   جدتي

الى من غرس الغرس و رعاه   والدي

إلى من دعت لي في اوقات المطر   والدتي

الى من ساعدني في مسيرتي الدراسية   خالتي

الى من شارف على تعليمي   اساتذتي

و الى كل من تمنى التوفيق لي

الشكر والتقدير

الحمد لله رب العالمين والصَّلاة والسَّلام على سيد
المرسلين و على اله
اجمعين

لابد لنا ونحن نخطو خطواتنا الاخيرة في الحياة الجامعية
من وقفة نعود إلى أعوام قضيناها في رحاب الجامعة مع
أساتذتنا الكرام

الذين قدموا لنا الكثير باذلين بذلك جهودا كبيرة في بناء
جيل الغد تبعث
الأمة من جديد

وقبل أن نمضي أقدم أسمى آيات الشكر والامتنان والتقدير
والمحبة الذين
حملوا أقدس رسالة في الحياة . . . .
إلى الذين مهدوا لنا طريق العلم والمعرفة .

إلى أساتذتنا الأفاضل وبالأخص الى الأستاذة (أميرة عبد
الواحد فنجان)

Abstract

The K-Means clustering algorithm is proposed by Mac Queen in 1967 which is a partition-based cluster analysis method. It is used widely in cluster analysis for that the K-means algorithm has higher efficiency and scalability and converges fast when dealing with large data sets. Precise clustering of computerized MRI brain Images presents the opportunity of early diagnosis for the different diseases such as the cancer . MRI images centers suffer from expert shortage, experiencing and high loads Particularly in developing countries. It prompted many researchers to make further improvements on the section is responsible for Medical image processing. Reducing the work area In medical image by separating the parts of MRI brain Image was a major point Which the radiologists targete.

المحتويات

Chapter One

Introduction of K-means clustering

1.1 Introduction

K-means clustering is a method of cluster analysis which aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean. It is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem [2] . K- means clustering is very useful in exploratory data analysis and data mining in any field of research, and as the growth in computer power has been followed by a growth in the occurrence of large data sets. Its ease of implementation, computational efficiency and low memory consumption has kept the k-means clustering very popular, even compared to other clustering techniques[1]. Such other clustering techniques include connectivity models like hierarchical clustering methods) [1]. These have the advantage of allowing for an unknown number of clusters to be searched for in the data, but are very costly computationally due to the fact that they are based on the dissimilarity matrix. Also included in cluster analysis methods are distribution models like expectation-maximization algorithms and density models.

The k-means algorithm is an evolutionary algorithm that gains its name from its method of operation. The algorithm clusters observations into k groups, where k is provided as an input parameter. It then assigns each observation to clusters based upon the observation’s proximity to the mean of the cluster. The cluster’s mean is then recomputed and the process begins again. Here’s how the algorithm works as shown in Fig 1:

2

1. The algorithm arbitrarily selects k points as the initial cluster centers (“means”).
2. Each point in the dataset is assigned to the closed cluster, based upon the Euclidean distance between each point and each cluster center.
3. Each cluster center is recomputed as the average of the points in that cluster.
4. Steps 2 and 3 repeat until the clusters converge. Convergence may be defined differently depending upon the implementation, but it normally means that either no observations change clusters when steps 2 and 3 are repeated or that the changes do not make a material difference in the definition of the clusters.

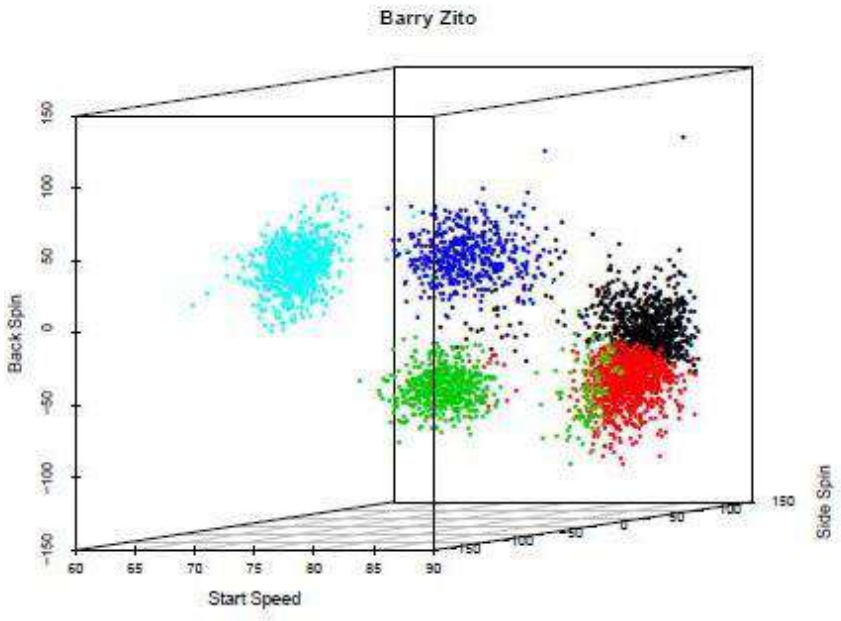


Fig. 1. K-means clustering algorithm

1.2 Motivation

The primary motivation behind studying K-means classification is its ability to classify unclassified data that give as flexibility in data analysis. As well as, we know how K-means classification is done to obtain a labeled dataset from an unlabeled dataset.

1.3 Applications of K-Means Clustering

K-Means clustering is used in a variety of examples or business cases in real life [2]،[3], like:

- Academic performance
- Diagnostic systems
- Search engines
- Wireless sensor networks
- Distance Measure

Academic Performance

Based on the scores, students are categorized into grades like A, B, or C.

Diagnostic systems

The medical profession uses k-means in creating smarter medical decision support systems, especially in the treatment of liver ailments.

Search engines

Clustering forms a backbone of search engines. When a search is performed, the search results need to be grouped, and the search engines very often use clustering to do this.

Wireless sensor networks

The clustering algorithm plays the role of finding the cluster heads, which collects all the data in its respective cluster.

Distance Measure

Distance measure determines the similarity between two elements and influences the shape of clusters.

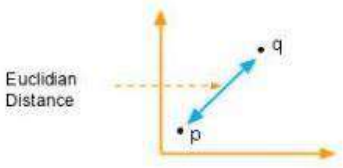
K-Means clustering supports various kinds of distance measures, such as:

- Euclidean distance measure
- Manhattan distance measure
- A squared euclidean distance measure
- Cosine distance measure

Euclidean Distance Measure

The most common case is determining the distance between two points. If we have a point P and point Q, the euclidean distance is an ordinary straight line. It is the distance between the two points in Euclidean space. The formula for distance between two points is shown below:

$$d=\sqrt{\sum_{i=1}^{n}(q_i-p_i)^2}$$



Squared Euclidean Distance Measure

This is identical to the Euclidean distance measurement but does not take the square root at the end. The formula is shown below:

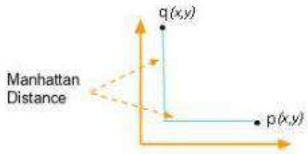
$$d=\sum_{i=1}^{n}(q_i-p_i)^2$$

Manhattan Distance Measure

The Manhattan distance is the simple sum of the horizontal and vertical components or the distance between two points measured along axes at right angles.

Note that we are taking the absolute value so that the negative values don't come into play. The formula is shown below:

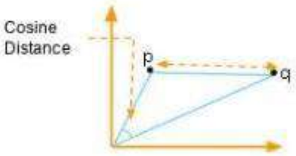
$$\mathbf{d}=\sum_{i=1}^{n} | q_x - p_x | + | q_{x} - p_y |$$



Cosine Distance Measure

In this case, we take the angle between the two vectors formed by joining the points from the origin. The formula is shown below:

$$\mathbf{d}=\frac{\sum_{i=0}^{n-1} q_i - p_x}{\sum_{i=0}^{n-1} (q_i)^2 \times \sum_{i=0}^{n-1} (p_i)^2}$$



1.4 Advantages and Disadvantages

Advantages

- Easy to implement
- With a large number of variables, K-Means may be computationally faster than hierarchical clustering (if K is small).
- k-Means may produce tighter clusters than hierarchical clustering
- An instance can change cluster (move to another cluster) when the centroids are recomputed.[2]

Disadvantages

- Difficult to predict the number of clusters (K-Value).
- Initial seeds have a strong impact on the final results.
- The order of the data has an impact on the final results.
- Sensitive to scale: rescaling your datasets (normalization or standardization) will completely change results. While this itself is

not bad, not realizing that you have to spend extra attention to scaling your data might be bad.[2]

Chapter Two

Medical Image Processing By K-means

2.1 K-Means Hard Clustering

Clustering is the process of grouping feature vectors into clusters (classes) in a self-organizing mode. K-Means is one of the most simplest unsupervised learning algorithms to solve clustering problem. K-Means is a hard clustering scheme that restricts each point of the dataset to exclusively just one segment (cluster) . K-Means technique is a pixel-based method, and it's complexity is relatively lower than other edge-based or region-based algorithms [4]. The procedure of its algorithm follows a simple and easy way to classify a given dataset through a predefined certain number of clusters. Occasionally the extracted features affect the clustering algorithm response [5]. For more details see. [6],[7]

2.2 Region Growing

Region growing is a pixel based segmentation. It depends upon selection initial points named seed points. Region growing algorithm involves investigating the neighboring pixels of seed points and figuring out which of the neighbors should be added to the region depending on predefined criteria such as intensity values, and this operation iterated until every pixel belongs to a discrete region. Region growing algorithm can correctly separate regions that have the same properties that predefined a prior and can provide the original images of clear edges with good segmentation results [3] . The basics and details of this algorithm can be found in Many researchers worked in medical image segmentation utilizing region growing such as:[3],[4],[8]

Experimental Datasets

In this work, six T2 and T1 MRI images of brain were adopted to implement the proposed technique on. Fig. (1) presents these images. Four of them are from local hospital in Baghdad city and the rest are from internet websites.[5]

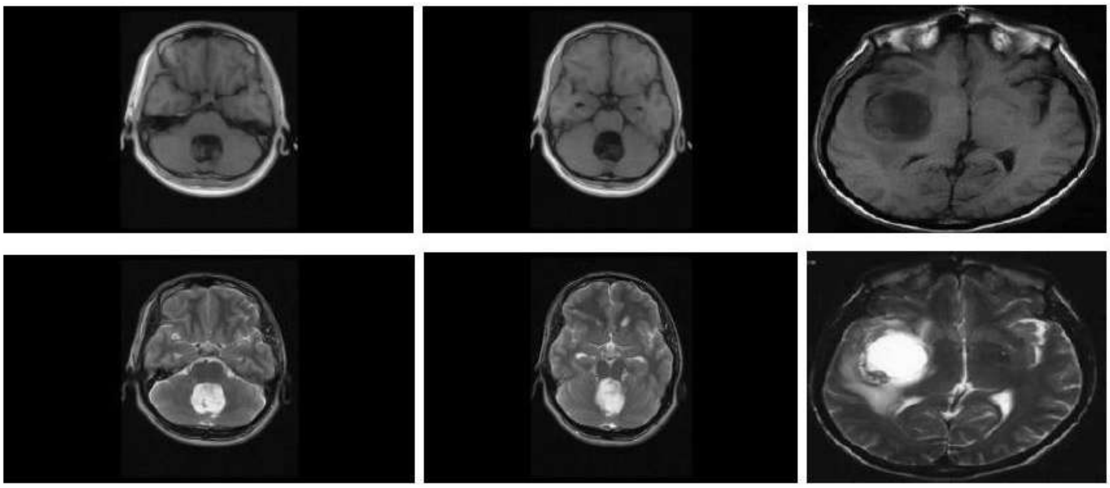


Fig.(1): The adopted experimental images, first line for T1 images while second line for their corresponding T2 images respectively.

The first two images of first and second lines are local images, while the third image of the two lines are website images. Local images need to disregard their background and the results of this process are shown in Fig.(2 a and b) for T1 and their corresponding T2 respectively.

a b

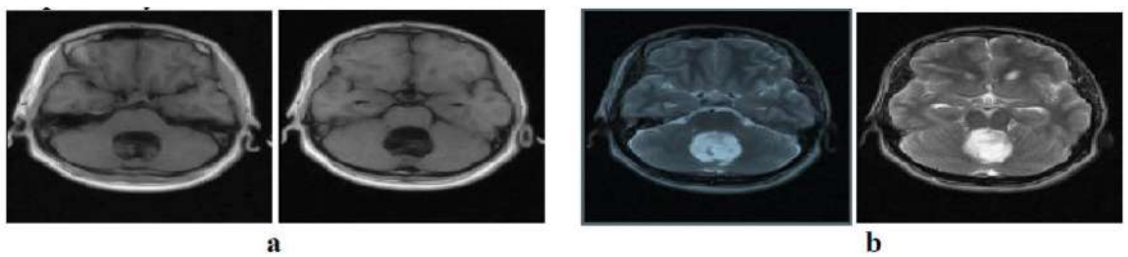


Fig.(2): Local experimental images after disregarding background. (a) for T1 images while (b) for their corresponding T2 images respectively.

Disregarding background process was achieved depending on the moment of the images, the steps of this technique was firstly proposed in our previous work.

2.3 Methodologies and Result

In this section, the results of implementing the proposed technique will be presented.

- **Region Growing Algorithm with Different Threshold**

In this step of the work, region growing algorithm was implemented, with different values of threshold, on T1 images and their corresponding T2 images to investigate its performance and to figure out the suitable values of the adopted threshold for the two experimental modalities' images [3],[7]. The results of this step for three T1 images are to Extract Tumor Regions from Region Growing Segmented Images [3]. In this step of work, an adaptive technique is proposed to extract tumor regions that were isolated by implementing region growing. This technique involves applying contouring process to separate the main regions that were resulted from region growing algorithm depending on their intensity and then selecting the tumor region only [7]. The results of this adaptive technique, for one T2 image, are shown in Fig. (3) for different threshold values: (25 35 45 and 65) as a sample.[1],[3],[5],[6]

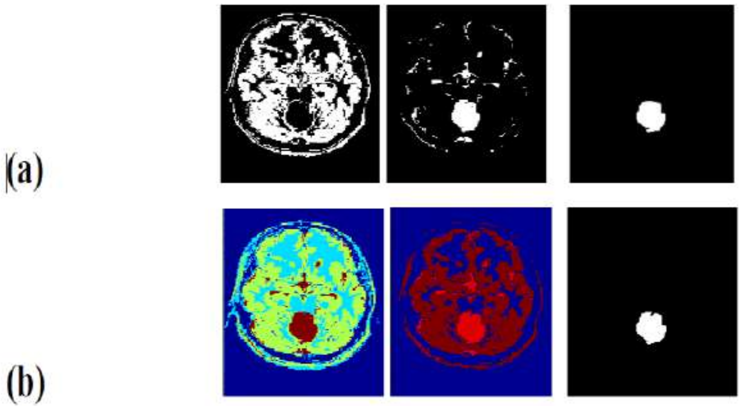


Fig (3): (a) K-Means segmented image and its four clusters, (b) Region growing based on K-Means

2.4 Conclusion.

K-means clustering is the unsupervised machine learning algorithm that is part of a much deep pool of data techniques and operations in the realm of Data Science. It is the fastest and most efficient algorithm to categorize data points into groups even when very little information is available about data.

References

1. L. Morissette, S. Chartier (2013). The k-means clustering technique: General considerations and implementation in Mathematical. Tutorials in Quantitative Methods for Psychology, Vol. 9(1), p. 15-24.

2. M. Santini (2016). Advantages & Disadvantages of k-Means and Hierarchical clustering (Unsupervised Learning). Uppsala University. P.4.

3. Wang, Zhiling, Andrea Guerriero, and Marco De Sario. 1996. “Comparison of Several Approaches for the Segmentation of Texture Images.” Pattern Recognition Letters 17 (5). Elsevier:509–21.

4. Gevers, Theo, and A W N Smeulders. 1997. “Combining Region Splitting and Edge Detection through Guided Delaunay Image Subdivision.” In Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference On, 1021–26.

5. Ali S. M., Loay K. Abood and Rabab Saadoon Abdoon, 2013, “Brain Tumor

Extraction In MRI images using Clustering and Morphological Operations Techniques", International journal of Geographical Information System Applications and Remote Sensing, Vol. 4, No. 1, pp: 12-25.

6. Law T. Y. and Heng P. A., 2000, "Automated extraction of bronchus from 3D CT Images of lung based on genetic algorithm and 3D region growing", Medical Imaging Proc. Of SPIE, Vol. 3979 , pp: 906-916.

7. Rabab Saadoon Abdoon, 2014, "Adaptive Techniques for Brain Tumor Detection", Ph.D. Thesis in image processing and remote sensing, University of Baghdad, Iraq.

8. Rabab Saadoon Abdoon, 2016, "Adaptive Technique Depending on Region Growing and Soft Clustering to Detect Tumors in Different Modalities of MRI Brain Images", Journal of Babylon University/ Pure and Applied Sciences, No.5, Vol. 24, pp:1333-1342.