LUNG IMAGE SEGMENTATION USING K-MEANS CLUSTERING ALGORITHM WITH NOVEL DISTANCE METRIC

D. Saraswathi¹, Dr. L. Mary Immaculate Sheela²

¹Ass.Professor, MCA Department, St.Xavier's College, Tirunelveli, Tamilnadu
²Research Supervisor, Dilla University, Ethiopia

Abstract - Clustering is one of the most important data mining techniques that can handle unlabeled data. K-means is a distance-based clustering algorithm. K-means groups the data objects into K disjoint clusters. K is a user specified parameter. Since, K-means is a simple algorithm it has been used in a wide variety of applications as well as in image processing. Its implementation is very simple and fast execution. The power of k-means algorithm is due to its computational efficiency and the nature of ease at which it can be used. Distance metrics are used to find similar data objects that lead to develop robust algorithms for the data mining functionalities such as classification and clustering. In this paper the K-Means clustering algorithm with various distance metrics are applied on images and the performance is analyzed. The various distance functions such as Euclidean distance, Manhattan distance, Chebyshev distance, Minkowski distance and Sorensen distance functions are used. In addition this paper proposes a pioneer distance metric with K-Means clustering algorithm to find better clusters for image segmentation. The work is implemented using MATLAB. The results are compared with the existing techniques by using various performance measures such as Precision Rate, Recall Rate, Sensitivity, Specificity and F-Measure. Experimental results indicate that the proposed distance metric with K-Means algorithm performs better in terms of performance evaluation metrics than the other distance metric functions.

Keywords - K-Means cluster, Euclidean distance, Manhattan distance, Chebyshev distance, Minkowski distance

I. INTRODUCTION

The process of separating an image into regions, or segments, is called segmentation. Segmentation is a widely studied area in computer vision. Researchers often try to optimise their segmentation algorithms to try and separate the objects in the image from the background.

There are different techniques for image segmentation and one of the most popular methods is k-means clustering algorithm. K-means clustering algorithm is an unsupervised algorithm and it is used to segment the interest area from the background. Clustering is an important data mining technique that has a wide range of applications in many areas like biology, medicine, market research and image analysis among others. Clustering is the process of partitioning a set of objects into different subsets such that the data in each subset are similar to each other. The similarity between various objects is defined by a distance measure. The distance measure plays an important role in obtaining correct clusters. The developing platform for the detection is MATLAB. Because it is easy to developed and execute. At the end, In section 2, we mention the related work; Section 3 describes overview of K Means algorithm; In section 4 various Distance Measures are defined; Section 5 contains Motivation and Justification of the work; Section 6 shows the Proposed Methodology; In section 7, Experimental Design and the results are examined; Section 8 contains the performance Evaluation and Section 9 concludes the paper.

II. RELATED WORK
Many distance measures have been proposed in literature for data clustering. Most often, these measures are metric functions; Euclidean distance, Manhattan distance, Minkowski distance, Hamming are such common functions. Jaccard index, Cosine Similarity and Dice Coefficient are also popular distance measures. For non-numeric datasets, special distance functions are proposed. For example, edit distance is a well known distance measure for text attributes.

There have been many works done in the area of image segmentation by using different methods. And many are done based on different application of image segmentation. K-means algorithm is the one of the simplest clustering algorithm and there are many methods implemented so far with different method to initialize the centre. And many researchers are also trying to produce new methods which are more efficient than the existing methods, and shows better segmented result. Some of the existing recent works are discussed here.

Dibya Jyoti Bora, Dr. Anil Kumar Gupta have done an experimental study using K-Means clustering algorithm in Matlab to cluster the iris and wine data sets with different distance measures and thereby observing the variation of the performance. In this experiment, they take “Cityblock”, “Euclidean”, “Cosine” and “Correlation”-these distance measurement techniques for distance calculations in the K-Means algorithm. Many comparisons have been done and they have found that city block distance shows better performance for both the datasets in terms of less computation time than Euclidean and cosine.

Richa Loohach & Kanwal Garg proposed the k-means clustering algorithm and various distance functions used in k-means clustering algorithm such as Euclidean distance function and Manhattan distance function. In this paper WEKA 3.6.5 version software for data mining is used and cpu.arff dataset is used for experimentation which can be obtained from UCI machine learning repository. From the experimental results they conclude that the number of iterations in the Euclidean distance function is generally more than the Manhattan distance function which shows that Manhattan distance function makes k-means algorithm less computational time complex than Euclidean distance function.

Madhu Yedla, Srinivasa Rao Pathakota, T. M. Srinivasa proposed Enhancing K-means clustering algorithm with improved initial center. A new method for finding the initial centroid is introduced and it provides an effective way of assigning the data points to suitable clusters with reduced time complexity. They proved their proposed algorithm has more accuracy with less computational time comparatively original k-means clustering algorithm. This algorithm does not require any additional input like threshold value. But this algorithm still initializes the number of cluster k and suggested determination of value of k as one of the future work.

K.A. Abdul Nazeer, M. P. Sebastian proposed an enhanced algorithm to improve the accuracy and efficiency of the k-means clustering algorithm. They present an enhanced K-means algorithm which combines a systematic method consisting two approaches. First one is finding the initial centroid and another is assigning the data point to the clusters. They have taken different initial centroid and tested execution time and accuracy. From the result it can be conclude that the proposed algorithm reduced the time complexity without sacrificing the accuracy of clusters.

All the previous works used Euclidean and Manhattan distance metric for calculation of the distances in K-Means clustering. None of the previous works had not employed another distance metrics such as Minkowski, Sorenson, Canberra and Chebyshev distance functions with K-means algorithm. This paper is aimed to analysis the impact of different distance metrics on K-means clustering algorithm for image segmentation.
III. K-MEANS OVERVIEW
K-means is a clustering algorithm, which partitions a data set into clusters according to some defined distance measure. Images are considered as one of the most important medium of conveying information. Understanding images and extracting the information from them such that the information can be used for other tasks is an important aspect of Machine learning. There are always K clusters. There is always at least one item in each cluster. The clusters are non-hierarchical and they do not overlap. Every member of a cluster is closer to its cluster than any other cluster because closeness does not always involve the ‘centre’ of clusters. K-means clustering in particular when using heuristics such as Lloyd's algorithm is rather easy to implement and apply even on large data sets. As such, it has been successfully used in various topics, ranging from market segmentation, computer vision and astronomy to agriculture. It often is used as a pre-processing step for other algorithms, for example to find a starting configuration. In statistics and data mining, 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. Figure 2 shows the flow chart of k-means algorithm which is relatively efficient and applicable only when mean is defined.

Algorithm:
1. Give the no of cluster value as k.
2. Randomly choose the k cluster centers
3. Calculate mean or center of the cluster
4. Calculate the distance between each pixel to each cluster center
5. If the distance is near to the center then move to that cluster.
6. Otherwise move to next cluster.
7. Re-estimate the center.

IV. DISTANCE MEASURES
Different measures of distance or similarity are convenient for different types of analysis. For numeric datasets, often used distance functions are Euclidean distance, Manhattan distance, Sorensen or Bray Curtis distance, Canberra distance and Chebyshev distance. Similarly for Boolean datasets and other non-numeric datasets other distance measures are used. Image distance is commonly used distance function for images and colour datasets. In the current paper, we study basic Euclidean distance, Sorensen or Bray Curtis distance, Manhattan distance, Chebyshev distance and Minkowski distance.

A **metric** on a set $X$ is a function (called the *distance function* or simply *distance*) $d : X \times X \to \mathbb{R}$ (where $\mathbb{R}$ is the set of real numbers). For all $x, y, z$ in $X$, this function is required to satisfy the following conditions:

1. $d(x, y) \geq 0$ (*non-negativity*, or separation axiom)
2. $d(x, y) = 0$ if and only if $x = y$ (*coincidence axiom*)
3. $d(x, y) = d(y, x)$ (*symmetry*)
4. $d(x, z) \leq d(x, y) + d(y, z)$ (*Triangle inequality*).

**Euclidean Distance**
The Euclidean distance or Euclidean metric is the ordinary distance between two points that one would measure with a ruler. It is the straight line distance between two points. In a plane with $p_1$ at $(x_1, y_1)$ and $p_2$ at $(x_2, y_2)$, it is $\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$. The distance is calculated using the formula

$$D_{XY} = \sqrt{\sum_{i=1}^{n}(x_{ik} - x_{jk})^2}$$

**Manhattan Distance**
Manhattan distance is a metric in which the distance between two points is the sum of the absolute differences of their Cartesian coordinates. In simple way of saying it is the absolute sum of difference between the x-coordinates and y-coordinates. Suppose we have two point A and B if we want to find the Manhattan distance between them, just we have to sum up the absolute x-axis and y-axis variation means we have to find how these two points A and B are varying in X-axis and Y-axis. In more mathematical way of saying Manhattan distance between two points measured along axes at right angles. It computes the absolute differences between coordinates of two objects:

$$Dist_{XY} = |X_{ik} - X_{jk}|$$

**Chebychev Distance**
Chebyshev distance is a variant of Minkowski distance where $p=\infty$ (taking a limit). Chebyshev Distance is also known as maximum value distance and is computed as the absolute magnitude of the differences between coordinate of a pair of objects.

$$Dist_{XY} = \max_k|X_{ik} - X_{jk}|$$

**Minkowski Distance**
The Minkowski distance is a generalized metric form of Euclidean distance and Manhattan distance.
Note that when \( p=2 \), the distance becomes the Euclidean distance. When \( p=1 \) it becomes city block distance. This distance can be used for both ordinal and quantitative variables.

**Sorensen Distance**

Sorensen distance is a normalization method that views the space as grid similar to the city block distance. Sorensen distance has a nice property that if all coordinates is positive; its value is between zero and one.

\[
\text{Dist}_{xy} = \frac{2 \| x_j - m_j \|}{\| x_j + m_j \|}
\]

**V. MOTIVATION AND JUSTIFICATION OF THE PROPOSED WORK**

The motivation is to devise a better segmentation method for medical images such as liver, brain, blood cell for detection of malignant tissue. Image segmentation has been identified as the key problem of medical image analysis and remains a popular and challenging area of research. Image segmentation is increasingly used in many clinical and research applications to analyse medical imaging datasets; which motivated us to present a snapshot of dynamically changing field of medical image segmentation.

Medical image segmentation is one of the most important tasks in many medical image applications, as well as one of the most difficult tasks. Medical image segmentation aims at partitioning a medical image into its constituent regions or objects, and isolating multiple anatomical parts of interest in the image. The accuracy of segmentation often determines the final success or failure of the entire application. If segmentation is done wrongly, the reconstruction will be erroneous. Therefore, considerable care should be taken to improve the reliability and accuracy of segmentation in medical image analyzing and processing.

We make use of K-means clustering algorithm, which is an unsupervised method, to provide us with a primary segmentation of the image. Distance metric is one of the key steps in the unsupervised learning process. Many parameters affect the final clustering results of K Means algorithm such as, algorithm initialization and distance metric. As changing the distance metric might change the final clustering results so, it is important to evaluate the impact of the distance metrics on K-means algorithm in clustering.

One of the demerits of K-Means algorithm is random selection of initial centroids of desired clusters. In this paper it was overcome by proposed K-Means with initial cluster centroid selection process for finding the initial centroids to avoid selecting centroids randomly and it produces distinct better results. The basic K-Means not provide better result for highly overlapped data in medical images such as Leukaemia images because of using Euclidean distance which strongly affect the cluster formation and yields wrong results. To overcome the problem the proposed methodology uses the Chebyshev and Sorensen distance with K-Means algorithm.

**VI. METHODOLOGY**

**Proposed Methodology**

In the proposed method, first, it determines the initial cluster centroids by using Chebyshev Distance based on the equation 1 which is given in the following algorithm. The Proposed K-Means algorithm
is improved by selecting the initial centroids manually instead of selecting centroids by randomly. It selects 'K' objects and each of which initially represents a cluster mean or centroids. For each of the remaining objects, an object is assigned to the cluster to which it is the most similar based on the distance between the object and the cluster mean. It then computes the new mean for each cluster. This process iterates until the criterion function converges. The Proposed K-Means algorithm is explained in very detailed manner in the below algorithm.

Algorithm : Proposed K-Means

Steps:
1. Using Chebyshev distance as a dissimilarity measure, compute the distance between every pair of all objects as follow.
   \[ d_{ij} = \max(|x_i - y_j|) \text{ where } i,j=1,\ldots,N \]  
   (1)
2. Calculate Iij to make an initial guess at the centres of the clusters
   \[ I_{ij} = \frac{d_{ij}}{\sum_{k=1}^{n} d_{ij}} \text{ where } i,j=1,\ldots,N \]  
   (2)
3. Calculate
   \[ \sum_{i=1}^{n} I_{ij}^2 \]  
   at each object and sort them in ascending order.
4. Select K objects having the minimum value as initial cluster centroids which are determined by the above equation. Arbitrarily choose k data points from D as initial centroids.
5. Calculate the distance between each data point and cluster centers using the sorensen and Chebyshev distance metric as follows and find the minimum distance between the two distance and consider as the final distance.
6. Data point is assigned to the cluster center whose distance from the cluster center is minimum of all the cluster centers.
7. New cluster center is calculated using:
\[ V_i = \left( \frac{1}{c_i} \right) \sum_{x_i} \]
where, ‘ci’ denotes the number of data points in ith cluster.

8. The distance between each data point and new obtained cluster centers is recalculated.

9. If no data point was reassigned then stop, otherwise repeat steps from 6 to 8.

VII. EXPERIMENTAL DESIGN AND RESULTS

We used different types of medical images for the analysis. Matlab is used to implement the proposed algorithm. In this work, Lung medical image is taken as the source image. The image in Figure shows the Lung information. K-Means algorithm with different distance metrics such as Manhattan, Euclidean, Chebychev, Sorenson, Minkowski and the proposed distance metric are applied on the source images and experimented and the results are interpreted. We compare the result of k-means algorithm with proposed algorithm and it is shown in the Fig.

Figure 3 (a) Input image of Lung (b) K-Means with Euclidean Distance Metric (c) K-Means with Minkowski Distance Metric (d) 10 K-Means with Sorenson Distance Metric (e) K-Means with Chebychev Distance Metric (f) K-Means with Proposed Distance Metric.

VIII. PERFORMANCE OF EVALUATION

Performance measures are essential to measure the possible benefits of fusion and also used to compare results obtained with different algorithms. The impact of Manhattan, Euclidean, Chebychev, Sorenson and Minkowski distances metrics with K-means algorithm have been evaluated in terms of performance metric and compared with the proposed method. These performance metrics are as following:
8.1 METRICS
The impact of proposed distance metric in K-means algorithm has been evaluated in terms of performance metrics. To evaluate the performance of the proposed method several performance metrics are available. This paper uses the Precision Rate, Recall Rate, Sensitivity, Specificity and F-Measure to analyses the performance.

1. Precision Rate
Precision is calculated as the fraction of correct objects among those that the algorithm believes belonging to the relevant class. It can be loosely equated to accuracy and it will roughly answer the question: “How many of the points in this cluster belong there/ correctly classified?”

\[
\text{Precision} = \frac{TP}{TP + FP}
\]

where TP = True Positive (Equivalent with Hits)
FP = False Positive (Equivalent with False Alarm)

2. Recall Rate
Recall is roughly answers the question. E.g. "Did all of the objects that belong in this cluster make it in?" In other words, recall is the fraction of actual objects that were identified.

\[
\text{Recall} = \frac{TP}{TP + FN}
\]

Where TP = True Positive (Equivalent with Hits)
FP = False Negative (Equivalent with Miss)

3. F-Measure
F-measure is the ratio of product of precision and recall to the sum of precision and recall. The f-measure can be calculated as,

\[
F_m = \frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision} + \text{Recall}}
\]

4. Specificity
Specificity measures the proportion of negatives which are correctly identified such as the percentage.

\[
\text{Specificity} = \frac{TN}{(FP + FN)}
\]

where, TN – True Negative (equivalent with correct rejection)
FP – False Positive (equivalent with false alarm)

5. Sensitivity
Sensitivity also called the true positive rate or the recall rate in some field’s measures the proportion of actual positives.

\[
\text{Sensitivity} = \frac{TP}{(TP + FN)}
\]

where, TP – True Positive (equivalent with hit)
FN – False Negative (equivalent with miss)

8.2 Analysis
To analysis the performance of the proposed system, it is compared with various techniques by using the performance metrics which are mentioned above. This is shown in the below table and graphs.
The following tabular column shows the comparison between different distance metrics and the proposed distance metric. The Precision Rate, Recall Rate, Sensitivity, Specificity and F-Measure are calculated for all the distance metrics and compared with the proposed one.

<table>
<thead>
<tr>
<th>Distance Methods</th>
<th>Precision Rate</th>
<th>Recall Rate</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manhattan</td>
<td>0.65</td>
<td>0.67</td>
<td>0.64</td>
<td>0.66</td>
<td>0.63</td>
</tr>
<tr>
<td>Minkowksi</td>
<td>0.71</td>
<td>0.73</td>
<td>0.69</td>
<td>0.71</td>
<td>0.68</td>
</tr>
<tr>
<td>Squared Euclidean</td>
<td>0.79</td>
<td>0.81</td>
<td>0.79</td>
<td>0.80</td>
<td>0.77</td>
</tr>
<tr>
<td>Sorensen</td>
<td>0.82</td>
<td>0.84</td>
<td>0.81</td>
<td>0.82</td>
<td>0.79</td>
</tr>
<tr>
<td>Chebychev</td>
<td>0.85</td>
<td>0.87</td>
<td>0.84</td>
<td>0.86</td>
<td>0.83</td>
</tr>
<tr>
<td>Proposed</td>
<td><strong>0.91</strong></td>
<td><strong>0.92</strong></td>
<td><strong>0.88</strong></td>
<td><strong>0.89</strong></td>
<td><strong>0.86</strong></td>
</tr>
</tbody>
</table>

Table 1. Performance Evaluation of K-Means with different distance metrics

(a) Precision Rate Analysis

(b) F-Measure Analysis

(c) Sensitivity Analysis

(d) Specificity

(e) Recall Rate
IX. CONCLUSION

In this work, we analysed K Means clustering algorithm with different distance measures are employed. We ran our experiments on different medical images. Several interpretations are made for the K Means algorithm and the distance measures based on the results. In k-means clustering algorithm different types of distance functions can be used to measure the distance between that is the closeness of two objects. Cluster overlapping in cluster formation does not yield accurate medical image segmentation. K means algorithm is a popular clustering algorithm applied widely, but the algorithm which selects k objects randomly from clusters as initial centroids cannot always give a good and stable clustering. In this paper, we have presented a new approach for image segmentation using improved K Means algorithm with a proposed distance metric. In the experiment Manhattan, Euclidean, Chebychev, Sorenson, Minkowski and the new proposed distance metric with K-means algorithm are experimented to see the effect of these distance functions on clustering. Experimental results show that selecting centroids by the proposed algorithm can lead to a better clustering. As a conclusion, the K - means, which is implemented using the proposed distance metric gives best result than the K - means based on other distance metrics such as Manhattan Euclidean, Chebychev, Sorenson, Minkowski distance metrics.

Future Enhancement

In future, we will consider another different distance measures for K-Means algorithm with respect to a big image dataset and perform a comparison among them, thereby, try to propose a good one for the Image Segmentation. Also, we will try to extend our study for different Segmentation algorithms for Lung Segmentation.

REFERENCES


