Jan - March © 2017 Transactions; 4(13): 509-520

International Journal of Research and Applications

ISSN (online): 2349-0020 ISSN (print): 2394-4544 http://www.ijraonline.com/

Research Article

Application Trends of Enhanced K-Means Clustering Algorithm

A. Avani [1] and Dr. N. Satyanarayana [2]

Corresponding Author:
aavniabit@yahoo.com

DOI:
http://dx.doi.org/
10.17812/IJRA.4.13(85)2017

Manuscript:
Received: 14th Jan, 2017
Accepted: 7th Mar, 2017
Published: 25th Mar, 2017

Publisher:
Global Science Publishing Group, USA
http://www.globalsciencepg.org/

ABSTRACT

In this paper, we present a novel algorithm for improving the performing of k-means clustering called enhanced k- means clustering. It organizes all the patterns in a k-dimensional structure such that one can find all the Datasets which are closest to a given dataset efficiently. The main intuition behind our approach is as follows. We have carried out the experimental analysis on various trends of applications namely facial recognition, medical images and social network analysis using enhanced k means clustering. Our experimental results demonstrate that our enhanced clustering algorithm can improve the Facial Recognition System recognition accuracy and also medical image enhancement, social network analysis of the direct k-means algorithm by an improved recognition accuracy of 2.6%.

Keywords: K-Means, Clustering Algorithm, Facial Recognition, Enhanced K-Means, Medical Image Enhancement, Social Network Analysis, Grimace face database, recognition.

[1] Research Scholar. Reg., No: PP.COMP.Sci & Eng. 0342C,
[1] Department of CSE, Rayalaseema University, Kurnool, Andhra Pradesh, India.
[2] Professor, Department of CSE, Nagole Institute of Technology, Hyderabad, India.

IJRA - Year of 2017 Transactions:
Month: January - March
Volume – 4, Issue – 13, Page No's:509-520
Subject Stream: Computers

Paper Communication: Author Direct

Paper Reference Id: IJRA-2017: 4(13)509-520

DOI: 10.17812/IJRA.4.13(85)2017

International Journal of Research and Applications
Jan - Mar © 2017 Transactions 4(13): 509-520

eISSN : 2349 – 0020
pISSN : 2394 – 4544
www.ijraonline.com

COMPUTERS RESEARCH ARTICLE

Application Trends of Enhanced K-Means Clustering Algorithm

A. Avani [1] and Dr. N. Satyanarayana [2]

[1] Research Scholar., Reg., No: PP.COMP.Sci & Eng. 0342C,
[1] Department of CSE, Rayalaseema University, Kurnool, Andhra Pradesh, India.
[2] Professor, Department of CSE, Nagole Institute of Technology, Hyderabad, India.
[1] aavniabit@yahoo.com and [2] nsn123@gmail.com

ABSTRACT

In this paper, we present a novel algorithm for improving the performing of k-means clustering called enhanced k- means clustering. It organizes all the patterns in a k-dimensional structure such that one can find all the Datasets which are closest to a given dataset efficiently. The main intuition behind our approach is as follows. We have carried out the experimental analysis on various trends of applications namely facial recognition, medical images and social network analysis using enhanced k means clustering. Our experimental results demonstrate that our enhanced clustering algorithm can improve the Facial Recognition System recognition accuracy and also medical image enhancement, social network analysis of the direct k-means algorithm by an improved recognition accuracy of 2.6%.

Keywords: K-Means, Clustering Algorithm, Facial Recognition, Enhanced K-Means, Medical Image Enhancement, Social Network Analysis, Grimace face database, recognition.

1. INTRODUCTION

Clustering is an important task for the discovery of community structures in networks. Its goal is to sort cases (people, things, events, etc.) into clusters so that the degree of association is relatively strong between members of the same cluster and relatively weak between members of different clusters. Merriam-Webster (2008) defined cluster analysis as a statistical classification technique for discovering whether the individuals of a population fall into different groups by making quantitative comparisons of multiple characteristics [1] [2]. Various clustering algorithms have been proposed in the literature in many different scientific disciplines such as Biometrics, Social Network Analysis and Medical Image Processing etc. Clustering algorithms are mainly categorized into several groups such as Partitioning methods, Hierarchical methods, Density based methods, and Grid based methods and Statistical method. Hierarchical clustering algorithms recursively find nested clusters either in agglomerative mode or in divisive mode [3] [4]. The most well-known hierarchical algorithms are single-link and complete-link; in single-link hierarchical clustering, the two clusters whose two closest members have the smallest distance are merged in each step; in complete-link case, the two clusters whose merger has the smallest diameter are merged in each step. Compared to hierarchical clustering algorithms, partitioned clustering algorithms find all the clusters simultaneously as a partition of the data and do not impose a hierarchical structure. The most popular and the simplest partition algorithm is K-means (Steinhaus, 1956). Berkhin (2009) [5].

In 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. This result in a partitioning of the data space into cells. The problem is computationally difficult (NP-hard), however there are efficient heuristic algorithms that are commonly employed and converge quickly to a local optimum. These are usually similar to the expectation maximization algorithm for mixtures of Gaussian distributions via an iterative refinement approach employed by both algorithms. Additionally, they both use cluster centres to model the data, however k-means clustering tends to find clusters of comparable spatial extent, while the expectation-maximization mechanism allows clusters to have different shapes [6][7].

Given a set of observations $(x_1, x_2, \ldots, x_n)$, where each observation is a d-dimensional real vector, k-means clustering aims to partition the n observations into k sets $(k \leq n)$ $S = \{S_1, S_2, \ldots, S_k\}$ so as to minimize the within-cluster sum of squares:

This non-hierarchical method initially takes the number of components of the population equal to the final required number of clusters. In this step itself the final required number of clusters is chosen such that the points are mutually farthest apart. Next, it examines each component in the population and assigns it to one of the clusters depending on the minimum distance. The centroid position is recalculated every time a component is added to the cluster and this continues until all the components are grouped into the final required number of clusters.

$$argmin \sum_{i=1}^{n} \sum_{x_i \in S_i}^{m} \|x_i - \mu_i\|^2 \quad (1)$$

Where μ_i is the mean of points in Si.

2. K MEANS ALGORITHM

Input: 'k', the number of clusters to be partitioned; 'n', the number of objects.

Output: A set of 'k' clusters based on given similarity function.

Steps:

i) Arbitrarily choose 'k' objects as the initial cluster centres';
ii) Repeat,

a. (Re) assign each object to the cluster to which the object is the most similar; based on the given similarity function;

b. Update the centroid (cluster means), i.e., calculate the mean value of the objects for each cluster;

iii) Until no change

Strengths of the K-Means algorithm:

1. Relatively scalable and efficient in processing large data sets; complexity is O (i k n), where i is the total number of iterations, k is the total number of clusters, and n is the total number of objects. Normally, k<<n and i<<n.
2. Easy to understand and implement.

3. ENHANCED K-MEANS ALGORITHM

This Enhanced K-Means algorithm has two phases. In the first phase, initially centroids are systematically found to produce clusters with better accuracy and in the second phase assigning the data points to the appropriate clusters. Because it provides a separate algorithm to calculate the initial centroids systematically, the final clusters are more accurate [8] [9].

Algorithm 3: Proposed Algorithm

Input: *k* is the number of clusters, *D= {d1, d2,.....,dn}* is set of n data items.

Output: A set of final *k* clusters.

Steps:

Phase 1: Determine the initial centroids of the clusters using Algorithm 1

Phase 2: Assign each data point to the appropriate clusters using Algorithm 2

At the first, compute the distance between each data point and all other data points in the data-point set *D*. Then locate the closest pair of data-points from the set *D* and create a new set of data points am having these two data point and delete them from *D*. Then find the closest data point to the set Am and add it to Am and delete it from *D*. Repeat this process until the number of data point in A_m reaches a threshold. When a threshold value met, perform step second and create a new data point set A_{m+1}. This process will be repeated for '*k*' such data point sets. Finally, the initial centroids will be obtained by averaging the vectors in each data point set [10].

Algorithm 1: Finding the initial centroids

Input: *D = {d1, d2... dn}* is set of *n* data items and *k* is the number of desired clusters

Output: A set of *k* initial centroids.

Steps:

1. Set *m* = 1;

2. First of all compute the distance between each data point and all other data- points in the set *D*;

3. Then find the closest pair of data points in the set *D* and create a new data-point set A_m which contains these two data- points, delete these two data points from the set *D*;

4. After that find the data point in *D* that is closest to the new data point set A_m, Add it to A_m and delete it from *D*;

5. Repeat step 4 until the number of data points in Am reaches *0.75*(n/k);*

6. If *m<k* and reaches a threshold, then at this point *m= m+1*, create another set of data by follow the steps from 3-5.

7. And now, for each data-point set A_m *(1<=m<=k)* find the arithmetic mean of the vectors of data points in Am, these means will be the initial centroids.

Steps for forming of final clusters are shown in Algorithm 4. Initially compute the distance between each data points and all centroids. Then assign the data points to the closest centroid and for each cluster, recalculate the centroids and repeat until the convergence criterion is met. At this point, final clusters are obtained. Euclidean distance is used to determine the distance between data points and cluster centroids [11].

Algorithm 2: Assigning data-points to clusters

Input: *D = {d1, d2,.....,dn}* is a set of n data items and *C= {c1, c2,.....,cn}* is set of *k* centroids

Output: A set of final *k* clusters.

Steps:

1. Firstly compute the distance between each data-point *di (1<=i<=n)* and all the centroids *cj*

(1<=j<=k) as *d (di, cj)*;

2. Now for each data-point *di*, find the closest centroid cj and assign *di* to cluster *j*.

3. Then for each cluster *j (1<=j<=k)*, recalculate the centroids;

4. Repeat

5. for each data-point *di*,

a. Compute its distance from the centroid of the present nearest cluster;

b. If this distance is less than or equal to the present nearest distance, the data-point stays in the cluster;

c. Else for every centroid *cj (1<=j<=k)* compute the distance *d (di, cj)*; End for;

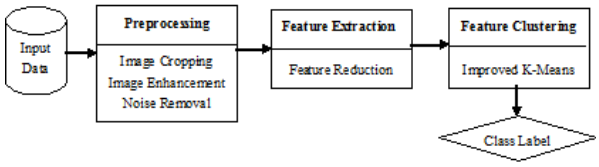
6. Assign the data-point di to the cluster with the nearest centroid *cj* End for (step (2));

7. For each cluster *j (1<=j<=k)*, Recalculate the centroids until the convergence criteria is met.

4. Face Recognition System with Enhanced K-Means Clustering Algorithms

The proposed Facial Recognition System divides the task of facial recognition into three major parts; pre-processing, facial feature extraction and classification. The step by step process for facial recognition is shown in figure1.

Pre-processing part includes four functions such as auto colour, auto brightness, and auto contrast and noise reduction. In facial feature extraction step face detection, segmentation by edge detection and feature extraction will be performed. At last, based on the extracted features Enhanced K-Means algorithm will be used in user identity classification [17] [18].



Architecture of the Proposed Face Recognition Approach

A. *Pre-Processing*

Pre-processing is the most important and the required step of the image processing. It is performed to get uniform and noise free image for further processing.

This step includes the following functions:

Auto Brightness- This function adjust the brightness of the image.

Auto Contrast- This function automatically calculates the favourable contrast for the image will increase the brightness of the image.

Auto Color- Auto color function adjust the color of the image.

Noise Reduction- Noise reduction will eliminate the unnecessary noise from the image.

B. *Feature extraction*

Facial feature extraction is the second foremost part of the face recognition system. This feature extraction part comprises of three phases: face boundary detection, segmentation by edge detection and feature extraction. Face boundary detection phase is performed to identify the face in the image that contains the eyes, nose and mouth. After that segmentation is performed to identify the region of interest. In this research work eyes, nose and mouth are taken as region of interest for the processing. Finally, the features of above regions are extracted.

a) Face Boundary Detection:

Face boundary detection phase is also a very important step for the face recognition. In this phase, the face boundary is detected and for that Successive Mean Quantization Transform (SMQT) features are used. The Successive Mean Quantization Technique performs an automatic structural breakdown of information. This information will be applied on local areas in an image to take out illumination insensitive features.

b) Segmentation by Edge Detection:

Segmentation of image means partitioning the image into multiple parts. In this system, segmentation is used to detect the interested regions such as eyes and mouth from images and for that edge detection method is used. After edge detection the region of interest is then cropped for feature extraction. Six edge detection methods are tested, named as: Roberts, Sobel, Prewitt, Laplacian of Gaussian, Zero-Cross and Canny. Canny method is chosen because it gives best results for edge detection.

c) Feature Extraction:

In this phase, the features of cropped interested region will be extracted and stored for classification. Eyes, Nose and mouth parts are taken as the region of interest and the features will be extracted. Two features will be calculated: first is density of pixels and second is ratio of height to width of cropped boundary regions. The calculation of ratio is done by dividing the face region into three zones: upper, middle and lower zones. Recognition accuracy of the system is depended upon the classification phase.

C. Face Classification:

Face recognition is the final step of the user identity recognition system. After feature extraction, facial features will be classified. For this, Enhanced K-Means algorithm is proposed [31].

Enhanced K-Means algorithm has two phases and in the first phase this will provide a separate algorithm for the calculation of initial centroids and in the second phase, assign the data points to the appropriate clusters. This will take calculated numerical values as an input and classify the user identity. In the existing system algorithm initial centroids are randomly selected whereas proposed algorithm calculates it systematically. In this work, Enhanced algorithm shows improvement in accuracy of the clusters and also efficient for huge dataset [27] [32].

5. ORL DATABASES

The ORL dataset was used to train the K-Means Clustering algorithm to generate the initial clusters. This dataset contains 40 individuals and each individual has 10 samples i.e., 400 samples. Samples are invariant to illumination and pose variations. Each image in the dataset was processed according to the training algorithm proposed above to generate mean, median and standard deviation of all the levels of wavelet decomposition, after addition of variable Gaussian noise. These values were then fed as input to K-Means Clustering algorithm with number of centroids fixed to 15. The K-Means Clustering algorithm was executed for 1000 loops to generate the final converged centroids. Thereafter the resultant centroids are used for denoising of sample test images in the wavelet domain using multilevel soft-threshold by varying the number of decomposition levels *(N)* and number of clusters *(K)*. It is observed from the above graphs that percentage of denoised increases with increasing the level of wavelet decomposition up to a certain extent. This rise in percentage of denoising is not linear with increase in number of levels of wavelet decomposition. It has been noticed that increasing the level of decomposition beyond 16 leads to reduction in percentage of denoised. A possible explanation for this outcome is that increase in the levels of wavelet decomposition leads to disintegration of the minute coarse details in the image which might not have been affected by the Gaussian noise added, and therefore once it is threshold, it results in loss of minute details which leads to a poor value of PSNR.

About Benchmark Databases

A. Grimace face database:

Grimace face database consists of 20 samples 18 individuals are there i.e., in total 360 images and resolution of each image is about 180x200 pixels. All the images of Grimace are minimum in illumination variation and maximum in expression variation. Background of all images is plain and having head, tilt and slant in translation.

Sample Images of Grimace face database

B. ORL face Database:

ORL face Database consists of 10 samples of 40 individuals in total 400 images are there with resolution about 92*112. The images of ORL are variant in Facial Expressions, illumination and in facial details (Glasses/ without Glasses). Translation in the frontal faces is too small.

Sample Images of ORL Face database

C. Yale face database:

Yale database consists of 11 samples for 15 individuals, i.e., in total 165 images are there. Resolution of each image is 320*243, samples are maximum variant in facial expression, details and also in illumination.

Sample Images of Yale Face database

6. MEDICAL IMAGE ENHANCEMENT WITH ENHANCED K-MEANS CLUSTERING

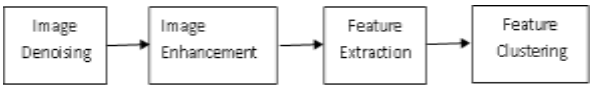
Clustering is defined as a method of organizing objects into teams whose members are similar in some way, so a cluster may be an assortment of objects that are similar between them and dissimilar to the objects and amalgamated to alternative clusters. The success of a cluster technique depends on the choice of a substantive intensity measure between the points that are clustered. There will be completely different clustering strategies based on the choice of similarity criterion that's used for bunch the objects [21] [23].

The medical space classification for disease analysis technique employed in this work makes use of clusters using enhanced based K-Means which is associate unsupervised technique of machine learning, works well for medical pictures. When distance between the information is employed it's known as distance-based cluster. If the cluster is based on any conception common to the objects it's known as abstract bunch. The selection of associate applicable similarity criterion can confirm the effectiveness of the clustering technique. Based on the approach the bunch is dispensed, there are completely different teams of enhanced based K means algorithms clustering. The two basic sorts are the stratified and partitioned cluster algorithms. Hierarchical cluster may be a sequence of partitions during which every partition is nested into the next partition within the sequence. Partitioned strategies of bunch are non-hierarchical since they generate one partition of the information in a trial to recover natural teams present within the information. Both the stratified and partitioned bunch strategies will be more divided into two types like agglomerated and dissentious bunch [24].

Agglomerated bunch starts with the different information objects placed in disjoint clusters and also the rule yield by merging the trivial clusters till one cluster results. Dissentious cluster performs the same task within the reverse order beginning with the complete set of information objects collectively cluster. Another technique of classifying the bunch algorithms is as serial and coincidental clustering. Serial procedures handle the information sets one by one whereas the coincidental method works with the entire data at a time [26].

The method that is adopted for the clustering in this work is a distance-based partitioned clustering. It uses Euclidian distance metric as the similarity criterion that guides the clustering process. Also it is a non-hierarchical method that does not find a hierarchical relation between the clusters formed. It is also a simultaneous clustering method whereas it considers the whole of the image at a time while performing the segmentation. Once the intensity analysis is over, at the next stage, the directional analysis over the image is performed using convolution filters. This is basically used to highlight the edge area so that the region based area separation will be performed. His segmentation is the combination of mathematical filters called convolution filter and morphological filters. After this stage, the complete region segmentation will be obtained. After the high level segmentation, enhanced k-means will be applied to improve the clustering outcome.

This segmentation process will compare the current cluster members with other cluster members and obtain the observation under two parameters called global member estimation and local member estimation. If some value or the intensity value satisfy the local membership, then it shows the existence of pixel or the component in particular cluster is valid. If the global membership satisfies, then the identification of the particular cluster will be done in which it actually satisfies the member. Based on this analysis, the component switching between the clusters will be done. At the final stage, the segmented area will be defined under some color model so that the colorization of different components over the image are performed. The architecture of the proposed research work is given in figure 3.



Recognition accuracy of Proposed Method.

Figure 3: Architecture of the proposed MRI Image Enhancement

Image De-noising: MRI images are usually corrupted by disturbances like Gaussian and Poisson noise. The vast majority of the de-noising algorithms assume additive white Gaussian noise. There are some algorithms that designed for Gaussian noise elimination, such as edge preserving bilateral filter, total variation, and non-local means. In this work, Median filtering is used as a nonlinear filter that is used as an effective method for removing noise while preserving edges. It works by moving pixel by pixel through the image, replacing each value with the median value of neighbouring pixels. The pattern of neighbours is called the ''window,'' which slides pixel by pixel over the entire image.

The median is calculated by first sorting all the pixel values from the window into numerical order, and then replacing the pixel being considered with the middle (median) pixel value. Image processing researchers commonly assert that median filtering is better than linear filtering for removing noise in the presence of edges. The output of this sub-step in pre-processing is the free noising MRI image.

In the scheme the below steps explores the implementation:

1. Filter the Brain Image in terms of brightness, contrast and image size.

2. Define the Number of clusters over the Image called N clusters

3. Implement Min Max analysis over image to identify frequency range.

4. Identify Intensity Variation=Range/N

5. For i=1 to N [Process all clusters] {

6. Obtain Center for Cluster (i) called Center(i) and the relative cluster variation}

7. Set AvgChange=MinThreshold

8. While (AvgChange<> ActualMod) {

9. Obtain the Segment Distance between pixel intensity and the cluster center

10. Update Variation min with minimum distance change using PSO

11. Include the pixels in different cluster based on intensity change

12. Take the Mean values to update the cluster centroids}

13. Return Cluster List;

However, the pseudo code s as under for ready reference and perusal: -

Step 1: Loading a gray scaled image Content Based Retrieval (Image)

/* Here Image is the actual medical image on which the content retrieval is performed*/

{

Image=To Gray(Image)

/*Check the image format and convert it to 8-bit grayscale image*/

Image=Adjust (Image)

/*pre-process the Image and perform the adjustment over the brightness and contrast*/

Step 2: Convert the image from RGB colour space to L*a*b* colour space

Unlike the RGB colour model, L*a*b* colour is designed to approximate human vision.

There is a complicated transformation between RGB and L*a*b*.

(L*, a*, b*) = T (R, G, B).

(R, G, B) = T' (L*, a*, b*).

Perform the Intensity Range Analysis called MinInt and MaxInt over the image.

Step 3: Undertake clustering analysis in the (a*, b*) colour space with the K-means algorithm In the L*a*b* colour space, each pixel has a properties or feature vector: (L*, a*, b*).

Like feature selection, L* feature is discarded. As a result, each pixel has a feature vector (a*, b*).

Applying the K-means algorithm to the image in the a*b* feature space where K = 3 (by applying the domain knowledge)

Perform the intensity based clustering over the Image and divide the area in small area segment under distance based analysis.

Step 4: Label every pixel in the image using the results from K-means Clustering (indicated by three different grey levels)

Step 5. Perform the convolution filter to highlight the edge points so that region classification will be performed

Step 6. Perform Region classification over the image.

Step 7. Obtain the clusters and take it as input set for PSO.

Step 8. Perform the cluster analysis in terms of frequency and variation analysis called velocity.

Step 9. Obtain the Local Best and Global Best analysis over each segment.

For i=1 to Length (Image)

{

Dist1=Feature (i)-Local Best

Dist2=Feature (i)-Global Best

If (Dist1<Dist2)

{

Identify the appropriate Cluster for Feature (i) and switch the clustering

Update the cluster information

Update the Local Best and Global Best Parameters

}}

Perform the Colorization on obtained classification contents over the image

}

7. Social Network Analysis with Enhanced K-Means Clustering

Text analytics defines a set of statistical, machine learning methods which helps to obtain high quality information from textual sources to serve research, business intelligence etc., it generally involves a procedure to derive patterns from structured data. Text mining includes tasks such as Text categorization, Text summarization, Entity relation modelling, Text clustering and Opinion mining [[28].

Text mining associates with information retrieval, lexical analysis and natural language processing (NLP) techniques to convert text into data for analysis purpose [29].

The working procedure of text analysis is as defined below:

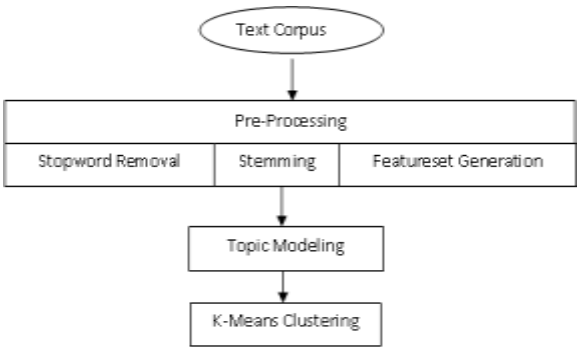
- The corpus consisting of textual material has to be identified and collected as a first step of analysis which is termed as Information Retrieval.
- Some text analytics approaches require statistical methods for performing analysis but several others systems require NLP techniques such as POS tagging, syntactic parsing for linguistic analysis.
- Named entity recognition involves certain statistical methods are used to identify named text features, abbreviations etc.,
- The noun phrases and other terms which are referring to single object.
- The associations existing among entities has to be identified and extracted.

Text data can be of structured format or unstructured format. Structured data is a well-organized data from which it gives a feasible scope of obtaining knowledge from it. Unstructured data does not have any pre-defined data model. It typically consists of text with dates, numbers and images etc., Due to presence of such features it leads to irregularities in text. In general, extracting useful knowledge from structured data is quite feasible using text mining techniques. When comes to unstructured data, a feasible process of obtaining intelligence from unstructured data is achieved by converting it into structured data. Unfortunately, all business data in current days is in unstructured format. Certain pre-processing steps of text mining are too performed efficiently to

obtain desired high quality information from unstructured data.

Typical quantitative data gathering strategies include:

- Experiments/clinical trials.
- Observing and recording well-defined events.
- Obtaining relevant data from management information systems.
- Administering surveys with closed-ended questions.



Architecture of the Proposed Opinion mining

Collecting Data: The data used for this research is a movie review dataset collected from twitter. There were about 25,000 tweets from the years 2003 to 2012.

Pre-processing: The pre-processing helps in finding the meaningful data, i.e., text mining, which includes the removal of unnecessary symbols, numbers and other if needed was done along with tokenization and stemming (remove the words with suffixes for example from the wanted "ed" is removed and it gives only "want". In the same way, coming, going, reading etc.).

The TF-IDF matrix was found.

The output from this phase is the data which don't contain any unnecessary tweets and spam tweets (meaningless tweets).

For opinion mining two functions are written called classify emotion () and Classify polarity (). The function classify emotion () deals with the genuine tweets obtained after the pre-processing phase and classify the emotion the tweet consists of Whether the person's emotion is anger, disgust, like, dislike etc.

TABLE I : RELATION AMONG EMOTION CLASSIFICATION AND POLARITY DEFINITION

Emotion Classification	Polarity Definition
Like, Love, Happy	Positive
Disguising, Sad, dislike	Negative

Now the tweets with the emotion and polarity were taken in a separate text file. A model was built using K-Means Clustering and further the accuracy is predicted. But in the review tweets the single tweet may consist of two or more number of opinions. So to predict those scores too, a word net lexicon dictionary which consists of a dictionary of positive and negative words were loaded. Later the opinion tweets which were stored in a separate file were loaded.

Now, by comparing the tweets with the lexicon dictionary the tweets carrying positive or negative opinions were easily recognized.

In the proposed work, we have used corpus which consists of movie review dataset (DB1) and twitter dataset (DB2) and analysis is performed by disjoining the corpus into training and testing tests. The cross validation technique divides the datasets into 10-fold in a random way. One division of the disjoined dataset is used for testing and the rest of the partitions for training.

The process is iterated 10 times, every division has dataset for testing and rest of parts as training sets. Six measures are used in the proposed work to evaluate the efficiency of opinion mining analysis. They are TP Rate, FP Rate, F-Score, Precision, Recall and Accuracy. When we select an algorithmic model for evaluation it is necessary to choose a metric for performance analysis. In case of two class problems, the test case could be as positive or negative.

In general, for a chunk of test cases, consider if TP represents number of opinions that the system accurately distinguishes as positive when reviews are labelled with positive. FN signifies the number of opinions that are negatively distinguished by the system when labelled as positive.

FP defines the number of opinions that are positively defined by the system when labelled as negative. TN specifies the number of opinions that the system accurately distinguishes as negative. The accuracy of a classification algorithm on test reviews is analysed as follows,

$$\text{Accuracy} = \frac{TN + TP}{TN + FP + FN + TP} \qquad (2)$$

The number of wrongly classified reviews to the overall review samples is termed as overall error

rate. Errors of Category-I define the number of reviews are classified as positive when labelled as negative. Errors of category-II define the number of reviews are classified as negative when labelled as positive.

$$\text{Overall error rate} = \frac{FP + FN}{number\ of\ samples} \quad (3)$$

$$\text{Category} - \text{I error rate} = \frac{FP}{number\ of\ postive\ reviews} \quad (4)$$

$$\text{Category} - \text{II error rate} = \frac{FN}{number\ of\ negative\ rveiews} \quad (5)$$

8. Experimental Results and Discussions

Performance evaluation of proposed approach with variant number of training samples against different levels of decomposition on ORL face database are given in Table 1.

TABLE II. Accuracy of the proposed method with different levels of decompositions

No. of samples used for training	8 levels of decomposition	16 levels of decomposition
2	95.1	95.7
3	95.8	96.4
4	97.1	97.6
5	97.6	98.3

The proposed method is evaluated on benchmark databases like ORL, Yale, FERET and Grimace with varied number of training samples.

TABEL III . Comparison of Accuracy for different database images

No of samples used for training	Accuracy (%)			
	ORL	Yale	FERET	Grimace
5	98.3	95	90	100
4	97.2	92.5	87.5	98.2
3	96.4	90.0	85	97.8
2	95.7	88.75	85	96.6

This proposed method has achieved 98.3% of recognition accuracy and it is 2.6% more than with 2 training samples. Likewise, experimentation is continued on other benchmark databases and achieved 95%, 90% and 100% recognition accuracies over Yale, FERET and Grimace databases respectively.

Experimental results are clearly evident that the proposed method outperforms existing methods with Enhanced K-Means clustering.

Further, proposed method is evaluated in the application areas like Medical Image Enhancement and Social Network Analysis.

To evaluate the performance of the proposed denoising and segmentation algorithms, some of the sample images are collected from the web and prepared the database. Database contains various MRI and X-ray images with different noise factors. To quantitatively evaluate the denoising algorithm, Root Mean Square Error (RMSE) and Structural Similarity Metrics (SSIM) are used.

5 Common images are considered as input images downloaded from Internet. Low resolution images are considered and resized to 512x512 images and applied with 3 methods they are wavelet denoising, edge enhancement and with k-means clustering. The values are tabulated with consideration of intermediate Peak Signal to Noise Ratio (PSNR) and RMSE values.

Firstly, Gaussian noise is added to the original image then PSNR and RMSE values are 17.6015 and 33.416 respectively. Further, the image was denoised with the wavelets and feature clustering and segmentation is done with Enhanced with K-Means then PSNR & RMSE values are 28.6335 and 9.4747 respectively. Likewise, further experimentation is done with the Salt and pepper noise then PSNR and RMSE values are 20.9392 and 22.9762 respectively. With the proposed approach, PSNR and RMSE are updated to 38.7386 and 2.9601. Further, performance evaluation of the proposed method is done with speckle and poison noises and then PSNR and RMSE values are presented in the tables 8 and 9.

TABLE IV. PSNR and RMSE values with Gaussian noise

Method	PSNR	RMSE	MSE
Gaussian Noise	17.6015	33.416	1138.4926
Wavelet denoising	18.3688	30.888	954.115
Proposed method	28.6335	9.4747	89.7693

TABLE V. PSNR AND RMSE VALUES WITH SALT AND PEPPER NOISE

Method	PSNR	RMSE	MSE
Salt and pepper Noise	20.9392	22.9762	527.9055
Wavelet denoising	21.3566	21.8983	479.5349
Proposed method	38.7386	2.9601	8.7623

TABLE VI. PSNR AND RMSE VALUES WITH POISON NOISE

Method	PSNR	RMSE	MSE
Poison Noise	32.0696	6.3791	40.6926
Wavelet denoising	33.8445	5.2001	27.0415
Proposed method	37.484	3.4201	11.6972

TABLE VII. PSNR AND RMSE VALUES WITH SPECKLE NOISE

Method	PSNR	RMSE	MSE
Speckle Noise	26.0682	12.7301	162.0598
Wavelet denoising	27.217	11.153	124.3892
Proposed method	33.461	5.4349	29.5376

These experimental results are clearly evident that the proposed Enhanced K-Means Clustering with Wavelet denoising enhances the image quality.

Further, we tested this enhanced k-means clustering algorithm on social network analysis to estimate the emotion accuracy in tweet message.

The performance of accuracy in recognition is tabulated in the following tables.

TABLE VIII. PERFORMANCE EVALUATION OF PROPOSED METHOD WITH 75 DIMENSIONS

Method	Accuracy on DB1	Accuracy on DB2
KNN	94.9%	93.9%
CHIRP	96.4%	96.1%
Naive Bayes	95.1%	96.1%
Enhanced K Means	97.7%	98.6%

TABLE IX. PERFORMANCE EVALUATION OF PROPOSED METHOD WITH 50 DIMENSIONS

Method	Accuracy on DB1	Accuracy on DB2
KNN	94.3%	95.2%
CHIRP	95.9%	96.9%
Naive Bayes	95.7%	96.7%
Enhanced K Means	97.4%	98.3%

9. CONCLUSION

In this paper, we have improved the k-means clustering algorithm by organizing all the patterns in a k-dimensional structure such that can find all the patterns which are closest to a given dataset efficiently. We performed on all the datasets which are potential datasets in closest at the root level. We used enhanced k means clustering on facial recognition, medical images and social network analysis. The proposed method is evaluated on benchmark databases like ORL, Yale, FERET and Grimace with varied number of training samples and achieved the good performance. Further, we explored the experiments on medical image enhancement and achieved the improvement in RMSE. We also studied and carried out experiments on the social network analysis using enhanced k-means algorithm and got the good accuracy in recognition of emotion in tweet messages.

REFERENCES

1) J. Zhu; H. Wang, "An improved K-means clustering algorithm, " 2010 The 2nd IEEE International Conference on Information Management and Engineering (ICIME), vol., no., pp.190, 192, 16-18 April 2010.
2) Huang, R. Harris 1993 "A comparison of several codebook generation approaches", IEEE Trans. Image Process. 2 (1), 108–112.
3) M. B. A. Daoud, S. A. Roberts, 1996. "New methods for the initialization of clusters", Pattern Recognition Lett. 17 (5), 451–45.
4) Likas, N. Vlassis, J. J. Verbeek, 2003. "The global K-means clustering algorithm", Pattern Recognition 36, 451–461.
5) S. S. Khan, A. Ahmad, 2004 "Cluster center initialization algorithm for k means clustering", Pattern Recognition Lett. 25 (11), 1293–1302.

6) M. Bishop, " Neural networks for pattern recognition ", Clarendon Press, Oxford, 1995.
7) Zhang, "Generalized k-harmonic means - Boosting in unsupervised learning", Technical Report HLP-2000-137", Hewlett-Packard Labs, 2000.
8) M. Fahim, A. M. Salem, F. A. Torkey and M. A. Ramadan, "An Efficient enhanced k-means clustering algorithm," journal of Zhejiang University, 10(7): 16261633, 2006.
9) K. A. Abdul Nazeer and M. P. Sebastian, "Improving the accuracy and efficiency of the k-means clustering algorithm," in International Conference on Data Mining and Knowledge Engineering (ICDMKE), Proceedings of the World Congress on Engineering (WCE-2009), Vol 1, July 2009, London, UK.
10) Chen Zhang and Shixiong Xia, " K-means Clustering Algorithm with Improved Initial center," in Second International Workshop on Knowledge Discovery and Data Mining (WKDD), pp. 790-792, 2009.
11) Yuan, Z. H. Meng, H. X. Zhangz, C. R. Dong, " A New Algorithm to Get the Initial Centroids," proceedings of the 3rd International Conference on Machine Learning and Cybernetics, pp. 26-29, August 2004.
12) Koheri Arai and Ali Ridho Barakbah, "Hierarchical K-means: an algorithm for Centroids initialization for k-means," department of information science and Electrical Engineering Politechnique in Surabaya, Faculty of Science and Engineering, Saga University, Vol. 36, No.1, 2007.
13) S. Deelers and S. Auwatanamongkol, "Enhancing K-Means Algorithm with Initial Cluster Centers Derived from Data Partitioning along the Data Axis with the Highest Variance," International Journal of Computer Science, Vol. 2, Number 4.
14) Mc Queen J, "Some methods for classification and analysis of multivariate observations," Proc. 5th Berkeley Symp. Math. Statist. Prob., (1): 281–297, 1967.
15) Bhattacharya and R. K. De, "Divisive Correlation Clustering Algorithm (DCCA) for grouping of genes: detecting varying patterns in expression profiles," bioinformatics, Vol. 24, pp. 1359-1366, 2008.
16) Margaret H Dunham, Data Mining-Introductory and Advanced Concepts, Pearson Education, 2006.
17) Elmasri, Navathe, Somayajulu, Gupta, Fundamentals of Database Systems, Pearson Education, First edition, 2006.
18) J. Pérez, R. Pazos , L. Cruz, G. Reyes, R. Basave, and H. Fraire "Improving the Efficiency of the K-means Clustering Algorithm Through a New Convergence Condition", Gervasi and M. Gavrilova (Eds.): ICCSA 2007, LNCS 4707, Part III, pp. 674–682. Springer-Verlag Berlin Heidelberg 2007.
19) Technologies, I. (2014). Face Image Retrieval Using Facial Attributes By, 5(2), 1622–1625.
20) Wang, Y., Yang, X., Wang, Y., & Wu, D. (2011). K-means based clustering on mobile usage for social network analysis purpose K-Means Based Clustering on Mobile Usage for Social Networ k Analysis Purpose, (June 2017).
21) M, M. I. C., P, P. R., & A, G. R. (2007). A Fuzzy Clustering Approach for Face Recognition Based on Face Feature Lines and Eigenvectors, (August).
22) Smith, N. (2009). K-Means Clustering For Detection of Human Faces in Databases by.
23) Contributions, O., Koh, H. C., & Tan, G. (n.d.). Data Mining Applications in Healthcare, 19(2), 64–72.
24) Naranjo-kairuz, C. (n.d.). Business Intelligence and Data Mining in chronic patients management : HL7 V3-based data model,1–67.
25) Wallace, B. C. (2012). Machine Learning in Health Informatics : Making Better use of Domain Experts, (August).
26) Computing, M., & Sekhon, A. (2017). Face Recognition Using K-Means and RBFN, 6(2), 137–141.
27) Hadid, A., & Pietik, M. (n.d.). Selecting Models from Videos for Appearance-Based Face Recognition.
28) Zhao, P., & Zhang, C. (2011). A new clustering method and its application in social networks. PATTERN RECOGNITION LETTERS, 32(15), 2109–2118. https://doi.org/10.1016/j.patrec.2011.06.008
29) Sahu, N. (2012). Analysis of Social Networking Sites Using K- Mean Clustering Algorithm, 88–92.

30) An adaptive K-Means Clustering Algorithm and its Application to Face Recognition. (2016), (July).
31) Otto, C., & Klare, B. (2015). An Efficient Approach for Clustering Face Images.
32) Pal, R., & Satsangi, C. S. (2016). Facial Expression Recognition Based on Basic Expressions and Intensities Using K-Means Clustering, 5(2), 2014–2017.
33) Agrawal, U. (2015). K-Means Clustering for Adaptive Wavelet Based Image Denoising K-Means Clustering for Adaptive Wavelet Based Image Denoising, (March), 5–9. https://doi.org/10.1109/ICACEA.2015.7164681