International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2015): 78.96 | Impact Factor (2015): 6.391

K- Means Clustering Algorithm Applications in Data Mining and Pattern Recognition

Huda Hamdan Ali[1], Lubna Emad Kadhum[2]

[1, 2]Assistant Lecturer, computer Engineering Techniques,Al-Imam Alkadhum College, Baghdad-Iraq

Abstract: ***Clustering is a process of partitioning a set of data (or objects)into a set of meaningful sub-classes, called clusters, help users understand the natural grouping or structure in adata set. Clustering has wide applications, inEconomic Science (especially market research), Document classification,Pattern Recognition, Spatial Data Analysis and Image Processing.This paper focuses on clustering in data mining and image processing. K-means algorithm is the chosen clustering algorithm to study in this work. The paper include: the algorithm and its implementation, how to use it in data mining application and also in pattern recognition.***

Keywords: k-means, clustering, data mining, pattern recognition

1. Introduction

The k-means algorithm is the most popular clustering tool used in scientific and industrial applications[1]. The k-means algorithm is best suited for data miningbecause of its efficiency in processing large data sets. Clustering is one of the well-known Data miningtechniques to find useful pattern from a data in a largedatabase (Fayyad, 1996)[2],However working only on numeric values limits its use indata mining because data sets in data mining often havecategorical values.

The Data mining approaches has been known to aid the process of detecting an intrusion in network environment. Thus, clustering algorithms were widely used for intrusion detection. M. Jianlian[3] introduced the application on intrusion detection based on K- means clustering algorithm. K-means is used for intrusion detection to detect unknown attack. Lei Li, et al [4] introduced a novel rule-based intrusion detection system using data mining. They proposed an improvement over priory algorithm by bringing the concept of length-decreasing support to detect intrusion It's also used in image processing generally and especially in pattern recognition applications programs to achieve some purpose like segmentation, where it gave an efficient solution for this important case.Alan Jose, S. Ravi and M.Sambath[5],proposed Brain Tumor Segmentation using *K*-means Clustering and Fuzzy*C*-means Algorithm and its area calculation. In the paper, they divide the process into three parts, pre-processing ofthe image, advanced *k*-means and fuzzy *c*-means and lastly the feature extraction. First pre-processing is implementedby using the filter where it improves the quality of the image. Then the proposed advance *K*-means algorithm is used, followed by Fuzzy *c*-means to cluster the image. Then the resulted segment image is used for the feature extractionfor the region of interest. They used MRI image for the analysis and calculate the size of the extracted tumor region inthe image.

2. Clustering[1]

The process of grouping a set of physical or abstract objects into classes of similar objectsis called clustering. A cluster is a collection of data objects that are similar to one anotherwithin the same cluster and are dissimilar to the objects in other clusters. A cluster of dataobjects can be treated collectively as one group and so may be considered as a formof datacompression. Although classification is an effective means for distinguishing groups orclasses of objects, it requires the oftencostly collectionand labeling of a large set of trainingtuples or patterns,which the classifier uses tomodel each group.Clustering is also called data segmentation in some applications because clustering partitions large data sets into groups according to their similarity.

A good clustering method will produce high qualityclusters whichthe intra-class (that is, intra-cluster) similarity is high,the inter-class similarity is low,the quality of a clustering result also depends on boththe similarity measure used by the method and itsimplementation,the quality of a clustering method is also measured byits ability to discover some or all of the hidden patterns,however, objective evaluation is problematic usuallydone by human / expert inspection.

In general, the major clustering methods can be classified into the following categories Partitioning methods, Hierarchical methods, Density-based methods, Grid-based methods, Model-based methods.

3. K-Means

The most well-known and commonly used partitioning methods are k-means.The k-means algorithm takes the input parameter, k, and partitions a set of n objects into k clusters so that the resulting intra cluster similarity is high but the inter cluster similarity is low. Cluster similarity is measured in regard to the mean value of the objects in a cluster, which can be viewed as the cluster's centroid or center of gravity."How does the k-means algorithm work?" The k-means algorithm proceeds as follows.

First, it randomly selects k of the objects, each of which initially represents a cluster mean or center. 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 [1].K-Means clustering is one of the

most famous clustering algorithms applied in different types of domains such as Biology and Zoology, Medicine and Psychiatry, Sociology and Criminology, Geology,Geography and Remote sensing, Pattern recognition and market research, and Education etc…[6].

3.1 Algorithm: k-means

The k-means algorithm for partitioning, where each cluster'scenter is represented by the mean value of the objects in the cluster.

Input:
k: the number of clusters,
D: a data set containing n objects.
Output: A set of k clusters.

Method:
1) Arbitrarily choose k objects from D as the initial cluster centers;
2) Repeat
3) (Re) assign each object to the cluster to which the object is the most similar,based on the mean value of the objects in the cluster;
4) Update the cluster means, i.e., calculate the mean value of the objects foreach cluster;
5) Until no change;

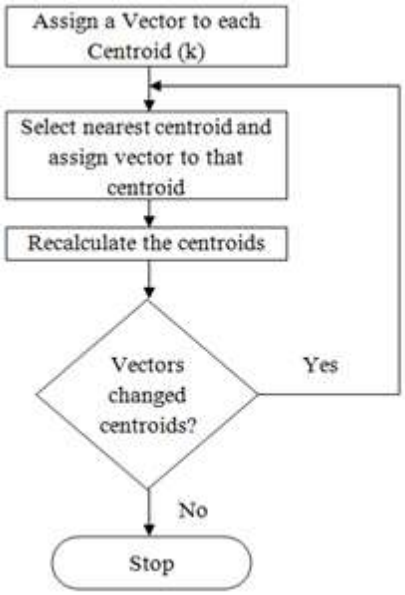
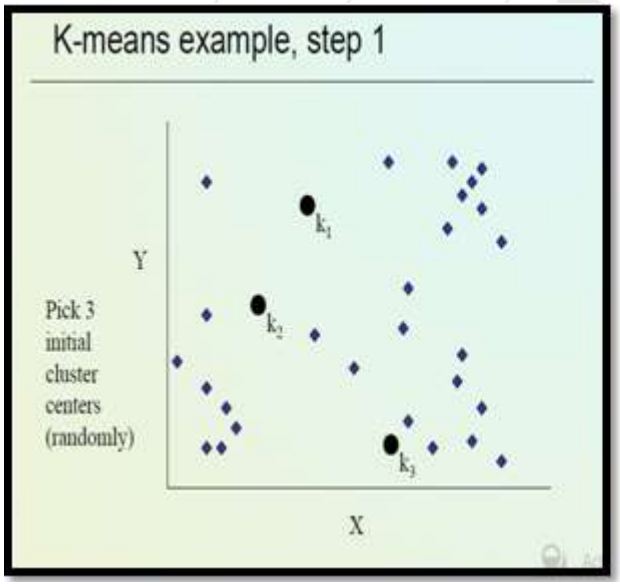
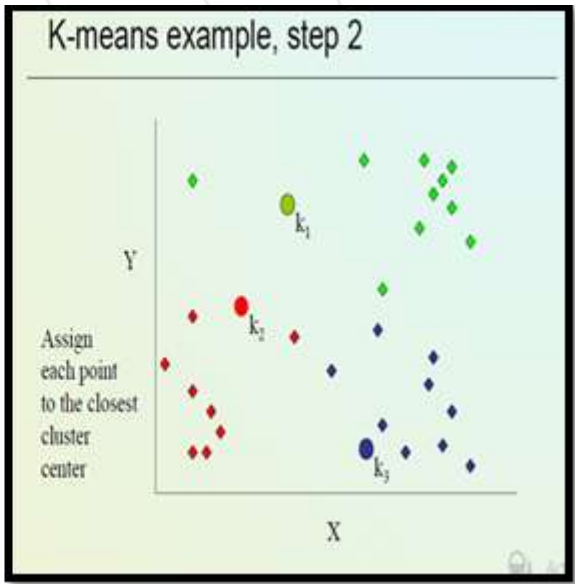


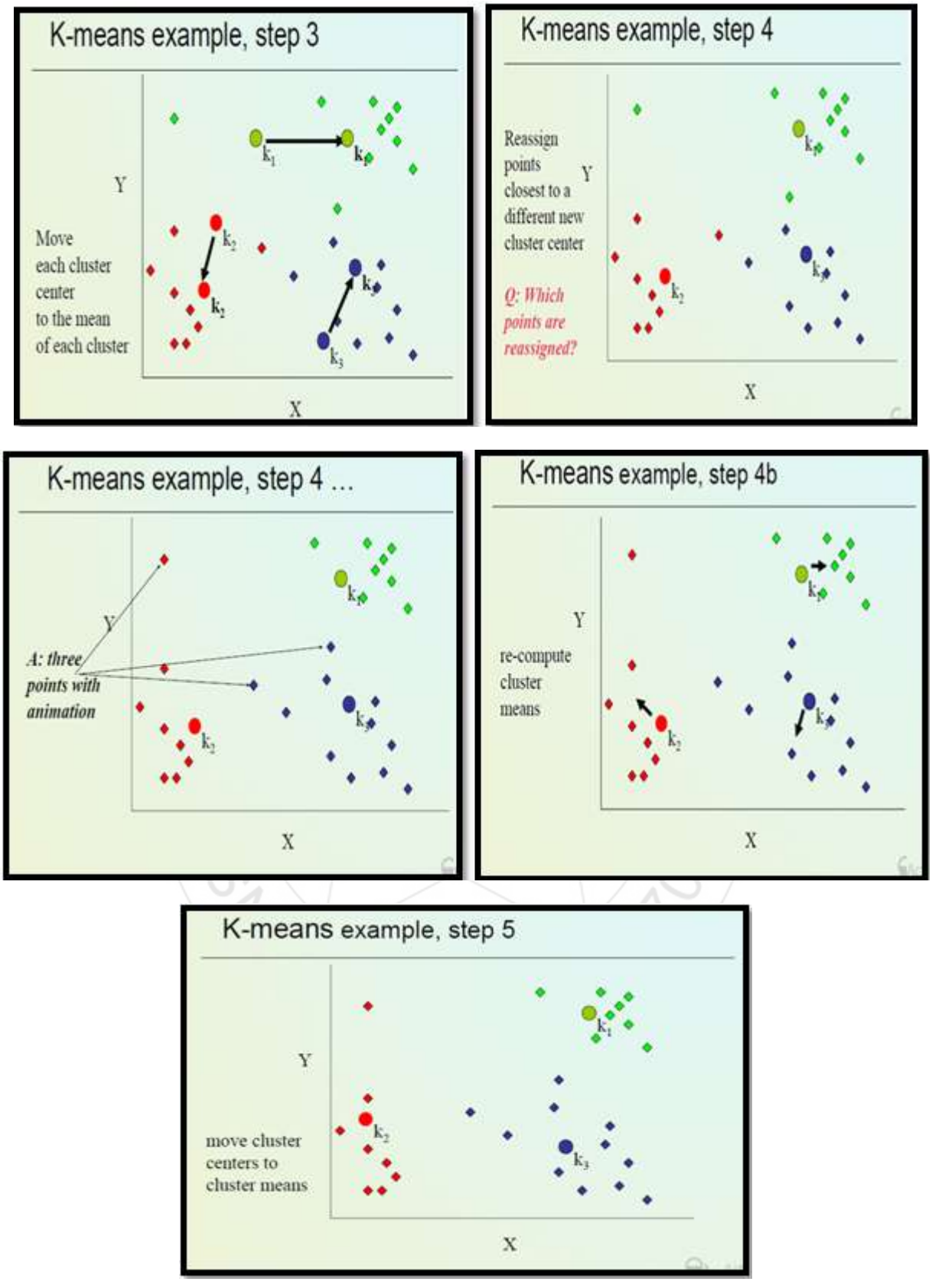
Figure 1: Flow chart of k-means algorithm

3.2 k-mean example

The below figures shows the steps of implementing k-means algorithm in details.







4. K-Mean Algorithm and Data Mining

The biggest advantage of the *k*-means algorithm in datamining applications is its efficiency in clustering largedata sets [7].Data mining adds to clustering the complications of very largedatasets with very many attributes of different types. This imposes uniquecomputational requirements on relevant clustering algorithms. A variety ofalgorithms have recently emerged that meet these requirements and weresuccessfully applied to real-life data mining problems. [8]

Today's business world is fast and dynamic in nature. It involves lot of data gathered from different sources.These data are stored in Data Warehouses. The mostchallenging

task of the business people is to transformthese data into useful information called knowledge. Datamining techniques are used to achieve this task. So to illustrate this case some examples are discussed in this paper.

4.1using of K-Means Clustering in Cloud Computing Environment[6]

Cloud Computing offers several benefits to the business organization to cut the initial investments to establish infrastructure for storage and compute. Many business organizations have already started migration of their business data into cloud data centers. Most often they need to mine useful information from the data stored in the cloud data centers with regards to business decisions. So the main objective of this work was to incorporate and implement K-Means Data mining technique into Cloud environment.For the experiment they took two data sets from the well-known real world data base "Machinelearning repository, 2012". One of data set they used was "Iris Dataset" (Fisher,1936). It consists of 5 attributes and150 instances. Theattributes are sepal width, sepal length, petal width, petallength and class label. It has three classes Iris flowers namely: C Iris setosa, C Iris versicolor,C Iris virginica. Theyhave created a database called "KMean" and twotables to store the data sets of iris and Blood TransfusionService Center Data Set in Cloud SQL using MySQL. Asone of the properties of K-Mean is that it assumes thenumber of clusters K is known in advance.

For the first experiment, they assume that k = 3.Figure 2 shows the output clusters of Iris dataset obtainedfrom the experiment. Since the value of K is 3 thenumbers of output clusters are three. The contents of eachcluster are displayed under the name cluster1, cluster2 andcluster3.

Figure 2: k-means action in cloud experiment on IRIS data set

So and after implementation of k- mean algorithm on that data set the result was:K-Means algorithm is more efficient algorithmfor mining large Databases and Cloud computing providessolution for storing largedatabase with less cost.

4.2 using K-Means Algorithm for EfficientCustomer Segmentation[9]

Data mining is the process of extracting meaningful information from dataset and presenting it in a human understandable format forthe purpose of decision support. The data mining techniquesintersect areas such as statistics, artificial intelligence,machine learning and database systems. The applications ofdata mining include but not limited to bioinformatics, weatherforecasting, fraud detection, financial analysis and customersegmentation. Customer segmentation is the subdivision of a businesscustomer base into groups called customer segments such thateach customer segment consists of customers who sharesimilar market characteristics. This segmentation is based onfactors that can directly or indirectly influence market orbusiness such as products preferences or expectations,locations, behaviors and so on. The importance of customer segmentation include, inter alia, the ability of a business tocustomize market programs that will be suitable for each of itscustomer segments; business decision support in terms ofrisky situation such as credit relationship with its customers.

Clustering has proven efficient in discovering subtle but tactical patterns or relationships buried within a repository of unlabeled datasets. This form of learning is classified under unsupervised learning. Clustering algorithms include k-Means algorithm, k-Nearest Neighbor algorithm, Self-Organizing Map (SOM) and so on. These algorithms, without any knowledge of the dataset beforehand, are capable of identifying clusters therein by repeated comparisons of the input patterns until the stable clusters in the training examples are achieved based on the clustering criterion or criteria. Each cluster contains data points that have very close similarities but differ considerably from data points of other clusters. Clustering has got immense applications in pattern recognition, image analysis, bioinformatics and so on. In this paper, the k-Means clustering algorithm has been applied in customer segmentation. A MATLAB program (Appendix) of the k-Means algorithm was developed, and the training was realized using z-score normalized two-feature dataset of 100 training patterns acquired from a retail business. After several iterations, four stable clusters or customer segments were identified. The two features considered in the clustering are the average amount of goods purchased by customer per month and the average number of customer visits per month. From the dataset, four customer clusters or segments were identified and labeled thus: High-Buyers-Regular-Visitors (HBRV), High-Buyers-Irregular-Visitors (HBIV), Low Buyers-Regular-Visitors (LBRV) and Low-Buyers-Irregular Visitors (LBIV). Furthermore, for any input pattern that was not in the training set, its cluster can be correctly extrapolated by normalizing it and computing its similarities from the cluster centroids associated with each of the clusters. It will hence be assigned to any of clusters with which it has the closest similarity.

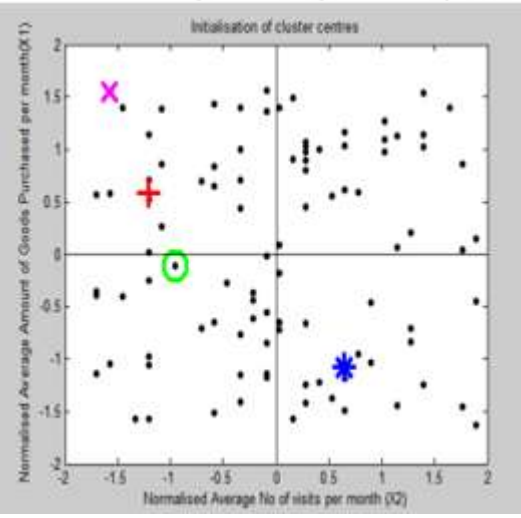


Figure 3: The initialization stage

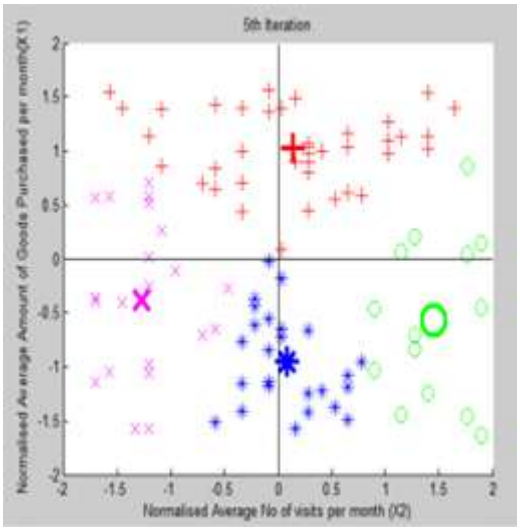


Figure 4: Positions of the centroids and their cluster members after of k-Means algorithm

The algorithm has a purity measure of 0.95 indicating 95% accurate segmentation of the customers. Insight into the business's customer segmentation will avail it with the following advantages: the ability of the business to customize market programs that will be suitable for each of its customer segments; business decision support in terms of risky situations such as credit relationship with its customers; identification of products associated with each segments and how to manage the forces of demand and supply; unraveling some latent dependencies and associations amongst customers, amongst products, or between customers and products which the business may not be aware of; ability to predict customer defection and which customers are most likely to defect; and raising further market research questions as well as providing directions to finding the solutions.

5. k-mean Algorithm in Pattern Recognition[10]

Extracting the information fromimages and understanding them such that the extractedinformation can be used for several tasks is an importantcharacteristic of Machine learning.Imagesegmentation is one of the initial steps in direction ofunderstanding images and then finds the different objects inthem.Image segmentation entails the separation or division of theimage into areas of similar attributes. It is one of the most crucial components of image analysis andpattern recognition and still is considered as most challengingtasks for the image processing and image analysis. It hasapplication in several areas like Analysis of Remotely SensedImage, Medical Science, Traffic System Monitoring, andFingerprint Recognition and so on.The segmentation methodincorporating clustering approaches encounters greatdifficulties when computing the number of clusters that arepresent in the feature space or extracting the appropriatefeature. This type of image segmentation is widely used due tothe simplicity of understanding and more accurate result.There are different methods 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.here some examples for using this algorithm in pattern recognition.

5.1 using k-mean algorithm to detect an infected part of fruit[11]

Plenty fruits are imported from theother nations such as oranges, apples etc. Manual identificationof defected fruit is very time consuming. anovel defect segmentation of fruits based on color features withK-means clustering unsupervised algorithm is presented.Colorimages of fruits for defect segmentation is used. Defect segmentation iscarried out into two stages. At first, the pixels are clustered basedon their color and spatial features, where the clustering process isaccomplished. Then the clustered blocks are merged to a specificnumber of regions. Using this two step procedure, it is possible toincrease the computational efficiency avoiding feature extractionfor every pixel in the image of fruits. Although the color is notcommonly used for defect segmentation, it produces a highdiscriminative power for different regions of image. Thisapproach thus provides a feasible robust solution for defectsegmentation of fruits. taken apple is taken as a case study andevaluated the proposed approach using defected apples. Theexperimental results clarify the effectiveness of proposedapproach to improve the defect segmentation quality in aspects ofprecision and computational time. The simulation results revealthat the proposed approach is promising.Image segmentation using k-means algorithm is quite usefulfor the image analysis. An important goal of imagesegmentation is to separate the object and background clearregardless the image has blur boundary. Defect segmentationof fruits can be seen as an instance of image segmentation inwhich number of segmentation is not clearly known.

The basic aim of the proposed approach is to segmentcolors automatically using the K-means clustering techniqueand L*a*b* color space. The introduced framework of defectsegmentation operates in six steps as follows:

Step 1. Read the input image of defected fruits.
Step 2. Transform Image from RGB to L*a*b* ColorSpace. We have used L*a*b* color spacebecause it consists of a luminosity layer in 'L'*channel and two chromaticity layer in 'a*' and'b*' channels. Using L*a*b* color space iscomputationally efficient because all of the colorinformation is present in the 'a*' and 'b*' layersonly.
Step 3. Classify Colors using K-Means Clustering in'a*b*' Space. To measure the difference betweentwo colors, Euclidean distance metric is used.
Step 4. Label Each Pixel in the Image from the Resultsof K-Means. For every pixel in our input, K-meanscomputes an index corresponding to acluster. Every pixel of the image will be labeledwith its cluster index.
Step 5. Generate Images that Segment the Input Imageby Color. We have to separate the pixels inimage by color using pixel labels, which willresult different images based on the number ofclusters.Programmatically determine the index of each clustercontaining the defected part of the fruit because K-means doesnot return the same cluster index value every time. But we cando this using thecenter value of clusters, which contains themean value of 'a*' and 'b*' for each cluster.

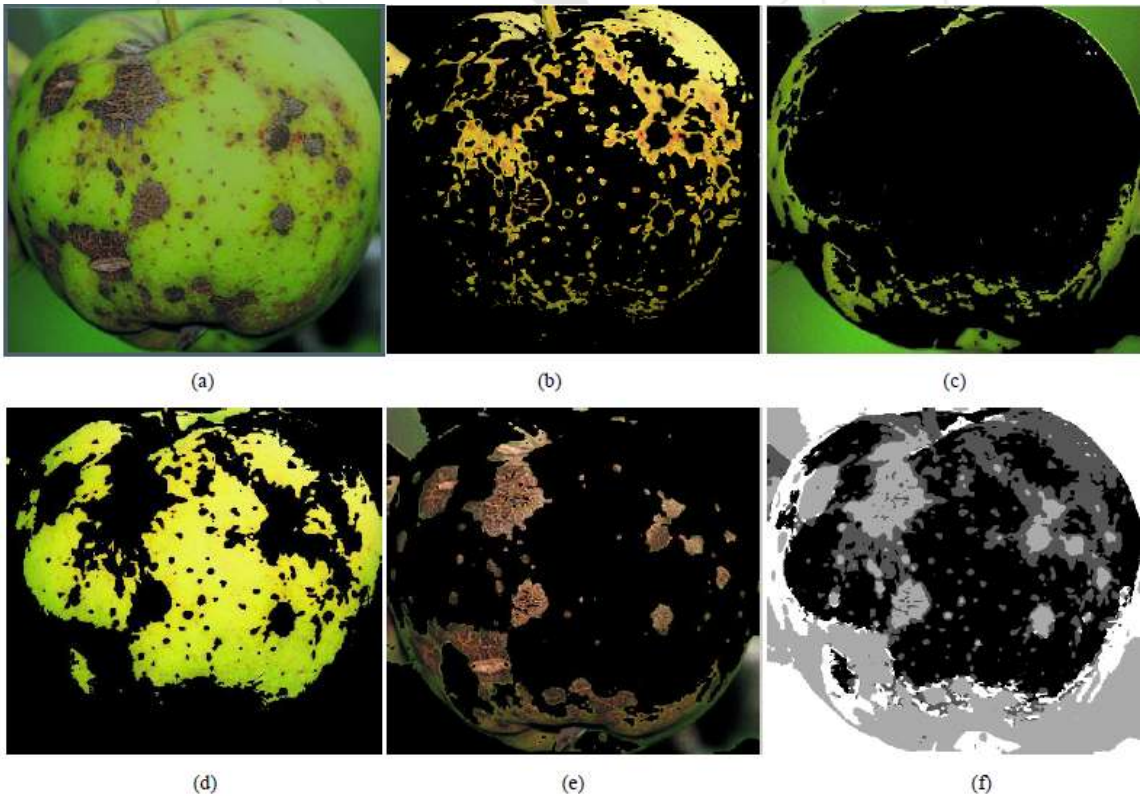


Figure 5: K-Means clustering for an apple fruit that is infected with apple scab disease with four clusters (a) The infected fruit image, (b) first cluster, (c) second cluster, (d) third cluster, and (e) fourth cluster, respectively, (f) single gray-scale image colored based on their cluster index.

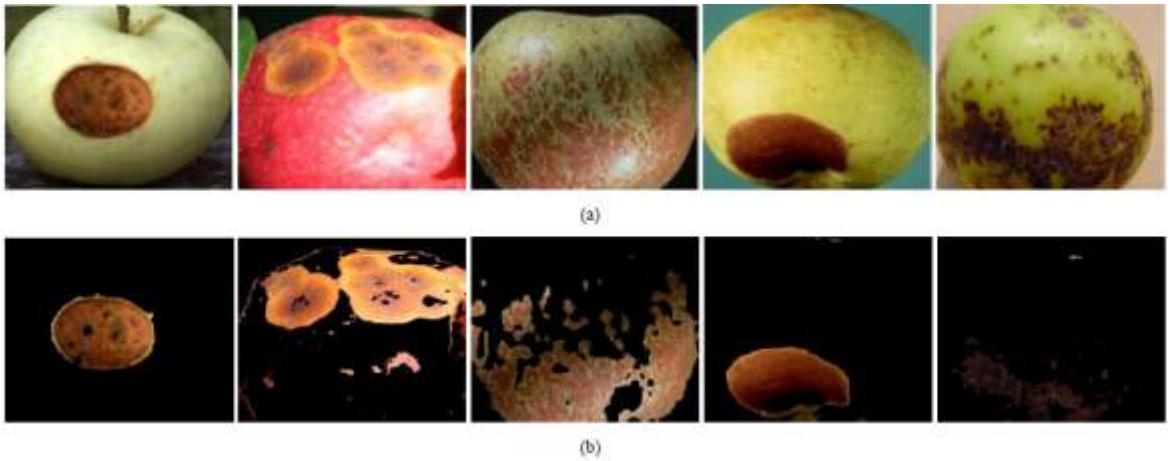


Figure 6: Defect segmentation results of apples (a) Images before segmentation, (b) Images after segmentation.

Experimental results suggest that the proposed approach isable toaccurately segment the defected area of fruits presentin the image. K-means based defect segmentation approach isalso segment defected area with the stem and calyx of thefruits.

5.2 using K-Means Clustering for Leukemia Image Segmentation[12]

During the unfolding measures that are taken for the purpose of leukemia detection, segmentation of blood cells is a vital step. There are several method can used for solving this stat .k-mean algorithm method is used and the results show that the segmentation based on K-means clustering gives good results.

The results show that thesegmentation using k-means clustering was implemented successfully as shown in fig.(7). The significant changes between segmented WBC (White Blood Cell) in and background in leukemia images can easily be seen. The fully segmented WBC (White Blood Cell) is achieved by application of the algorithm. And better results were obtained when K-Means Clustering was used compering with another algorithm result.

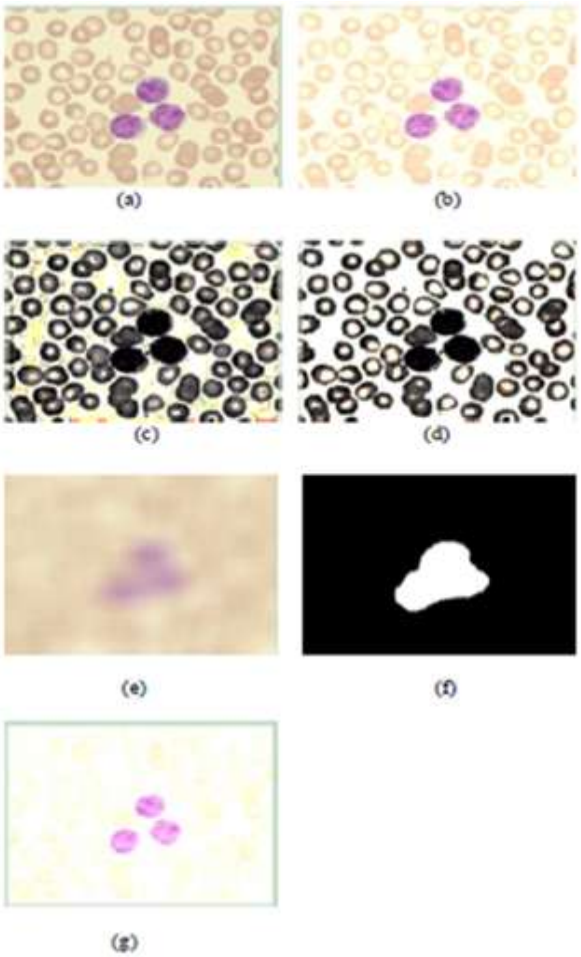


Figure 7: (a) Original image and resultant image after applying : (b) Linear Contrast, (c) HSI color Model, (d) K-Means Clustering, (e) Median Filter (f) Binary image and (g) Final segmented image.

6. Advantage and disadvantage of k-mean algorithm

6.1 Advantages

This clustering methodology which we described earlier, has some benefits comparing to others. The most important ones are:

- Lots of Applications – It has several live world implementations on many different subjects. We talk about the more relevant later in this article.
- Fast – Achieves the final result of its iterations in a fast way due to the simplicity of the algorithm.
- Simple and reliable – The process is fairly simple and always terminates, solving the problem with a solution set even for large data sets of information.
- Efficient – This method presents a good solution with relative low computational complexity for the clustering problem.
- Good Solutions – Provides the best result set specifically when data points are fairly separated.

6.2 Disadvantages

However this implementation has some problems which need to be addressed. We provide you a list of the major ones:

- No Categorical Data – One of the bigger problems of k-means clustering is that it can't be used on data entries that can't simulate a mean function.
- Set Number of Clusters – In this algorithm the number of partitions must be pre-defined. If this number is badly set, the implementation and results will suffer a lot. Therefore you should use techniques to estimate the number of clusters.
- Result Set – As we explained before, the result set of this process might not be optimal.
- Initialization Method – Depending on the chosen initialization process, the results will differ.

7. Conclusion

By studying several application of k-means algorithm in both side data mining and pattern recognition where its use as clustering method with data mining giving a promised results and using as segmentation result when it used in pattern recognition and also give a very good result ,so that's mean it's an efficient algorithm in both state .

References

[1] John A. Hartigan , *Clustering Algorithms*, John Wiley & Sons New York , London , Sydney , Toronto,1975 .
[2] Fayyad, U.M., G. Piatetsky Shapiro, P. Smyth And R. Uthurusamy, *Advances In Knowledge Discovery And Data Mining*,Aaai Press/The Mit Press, Pp: 573-592, 1996.
[3] M.Jianliang, S.Haikun And B.Ling, *"The Application On Intrusion Detection Based On K-Means Cluster Algorithm*", IEEE International Conference On Information Technology And Applications, 2009.
[4] Lei Li, De-Zhang Yang, Fang-Cheng Shen,*A Novel Rule-Based Intrusion Detection System Using Data Mining*,978-1-4244-5539, IEEE,2010.
[5] Alan Jose, S. Ravi And M. Sambath, Brain Tumor Segmentation Using *K*-Means Clustering And Fuzzy *C*-Means Algorithm And Its Area Calculation. In *International Journal Of Innovative Research In Computer And Communication Engineering*, Vol. 2, Issue 2, March ,2014.
[6] A. Mahendiran, N. Saravanan, N. Venkata Subramanian And N. Sairamm, *Implementation Of K-Means Clustering In Cloud Computing Environment*, Research Journal Of Applied Sciences, Engineering And Technology 4(10): 1391-1394, 2012.
[7] Zhexue Huang ,*A Fast Clustering Algorithm To Cluster Very Large Categorical Data Sets InData Mining*, Cooperative Research Centre For Advanced Computational Systems Csiro Mathematical And Information Sciences, Established Under The Australian Government's Cooperative Research Centers Program.
[8] RuiXu,*Survey Of Clustering Algorithms*, IEEE Transactions On Neural Networks, Vol. 16, No. 3, May 2005 .
[9] Chinedu Pascal Ezenkwu, Simeon Ozuomba, Constance Kalu, *Application Of K-Means Algorithm For Efficient Customer Segmentation: A Strategy For Targeted Customer Services*, (Ijarai) International Journal Of Advanced Research In Artificial Intelligence, Vol. 4, No.10, 2015.
[10] B.S.Vamsi Krishna, P.Satheesh, Suneel Kumar R., "*Comparative Study Of K-Means And Bisecting K-Means Techniques In Wordnet Based Document Clustering*", International Journal Of Engineering And Advanced Technology, Volume-1, Issue-6, August 2012
[11] Shiv Ram Dubey1, Pushkar Dixit2, Nishant Singh3, Jay Prakash Gupta4,*Infected Fruit Part Detection Using K-Means Clustering Segmentation Technique,International Journal Of Artificial Intelligence And Interactive Multimedia, Vol. 2, Nº 2.*,
[12] MashiatFatma, Jaya Sharma,*Leukemia Image Segmentation Using K-Means Clustering And Hsi Color Image Segmentation,International Journal Of Computer Applications (0975 – 8887) Volume 94 – No 12, May 2014.*