Breast Cancer Diagnosis Using K-means Methodology

Noor Kadhim Ayoob
Science College For Women
Noor.kadhum@gmail.com

Abstract

After lung cancer, breast cancer is the second cause of death among women. Due to the seriousness of the disease, research has stepped up to help diagnose this disease by providing medical personnel with a classification based computer systems that determine whether the patient is infected. This research focuses on the use method (K-means) for the diagnosis of breast cancer based on a global database known as (WBCD) dedicated to this purpose. The proposed method has proved its effectiveness in classification and the accuracy of the system is equal to 96.4861%.

Keywords: WBCD ; K-means; Breast cancer diagnosis; classification.

الخلاصة

يعد سرطان الثدي المسبب الثاني للوفاة بين النساء بعد سرطان الرئة, و نظراً لخطورة هذا المرض، انبرت البحوث للمساعدة في تشخيص هذا المرض عن طريق تجهيز الطاقم الطبي بأنظمة تصنيف تعتمد على الكمبيوتر لتحديد ما إذا كان المريض مصاب أم لا. يركز هذا البحث على طريقة استخدام (K-means) لتشخيص سرطان الثدي بالاعتماد على قاعدة بيانات عالمية معروفة باسم (WBCD) مخصصة لهذا النوع من الأبحاث. أثبتت هذه الطريقة فاعليتها في التصنيف وقد وصلت دقة النظام الى96.4861٪.

الكلمات المفتاحية : تصنيف, تشخيص سرطان الثدي, (K) من المتوسطات , قاعدة البيانات (WBCD)

1. Introduction

The goal of the classification is to assign a class to find previously unseen records as accurately as possible. Researcher has a wide range of options when it comes to classification, these options are including k-nearest neighbor, decision tree, neural networks, fuzzy logic techniques and genetic algorithm or even can merge more than one of these methods in the hope of getting the more powerful ,accurate classifier.

There are two types of classification, supervised and unsupervised. In the recent years, clustering is the process of gathering data items into groups (or clusters) using unsupervised classification [Thakare, 2015]. Items that belong to a particular cluster should be similar as much as possible and at the same time, they differ from the items that belong to other clusters.

K-means clustering is one of the simplest partition-based cluster analysis method [Nikhil, 2013]. The algorithm begins to specify the number of clusters (k) and the allocation of initial values of the centers, these values could be random or you can choose the first K items. By using this method as a classification tool, k-means can serve in the diagnosis of disease.

The rest of this paper is organized as following: after presenting important researches related to WBCD classification in section 2, section 3 provides a review of K-means method. The description of WBCD problem is given in section 4. Subsequently, the proposed system is discussed in Section 5. In section 6, the experimental results of the proposed system are shown. Finally, section 7 gives the conclusion and suggestions for future work.

2. Related work

` Because of the seriousness of this disease, many of the researches appeared in the diagnosis of breast cancer over the years and they are still ongoing research in this area.

In 2009, Karabatak and Cevdet built system that automatically gave the diagnosis of breast cancer based on the association rules and neural networks. The rules are invested to reduce the inputs of network by identifying important symptoms, the proposed method reduces the symptoms from 9 to 4 only to be adopted in multilayer Perceptron consists of four entry and made up a hidden layer of eleven cell with a single cell cells in the output layer. The research was conducted a number of experiments to study the effect of a number of symptoms used to diagnose the disease, when the network was run using the nine symptoms, the rate was 95.2% . After reducing the symptoms into only four, performance ratio was 95.6% [Karabatak, 2009].

In 2013, Mahau Nandi used MATLAB to experience three techniques for the diagnosis of cancer, a SVM (Supprt Vector Machine), neural networks , and the k-means. The researcher explained that the performance of (SVM) was the best where it has reached (96.7%). The ability of classification of neural network was (95.85%) while the performance of (k-means) was low compared with its counterparts, it did not exceed (94.13%) [Mahau, 2013].

In the same year, Indira Muhic research to study the possibility of diagnosing cancer using a technique Fuzzy C-Means(FCM), which distributed database instances into two groups (clusters) where each case at the database belongs to both clusters by certain probability. The results showed that the application of this technique is able to identify benign cases by 100%, while the system is recognized 87% of infected cases which means that the accuracy of the system is 93.5% [Indira, 2013].

3. K-means technique

Clustering is an iterative process [Bhagwati Charan Patel ,2010] of gathering items into groups ,called clusters, so that the elements in a particular group are identical with each other, and at the same time be different as possible from the elements in the other groups [Nikita Ghiya and etal,2015]. Because this method working on partioning of large databases into groups, it is sometimes called (data segmentation). This method can work without the need for prior information and this is contrast with supervised classification [Wanli Xiang ,2015] .

For databases of m instances and n features, the method starts by specifying the number of clusters and choose centers, in the simplest cases, this is done at random or choose the first k instances. Then all points (instances) allocated to the closest cluster by calculating the distance between the instances and the centers of the clusters. Clustering methods commonly used Euclidean distance to judge the similarity of the two data vectors [Dharmendra,2014]. The cluster center is updated based on the points belonging to this cluster. When all points desist from changing the process will stop [Sridevi ,2014]. Figure 1 is shown the phases of k-means for clustering randomly generated data into three classes

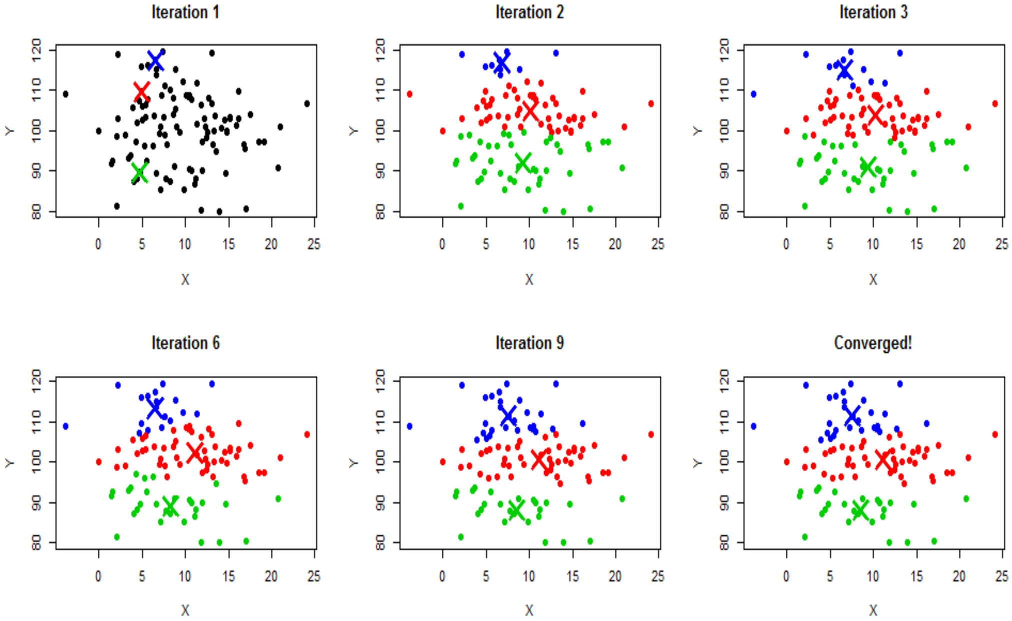
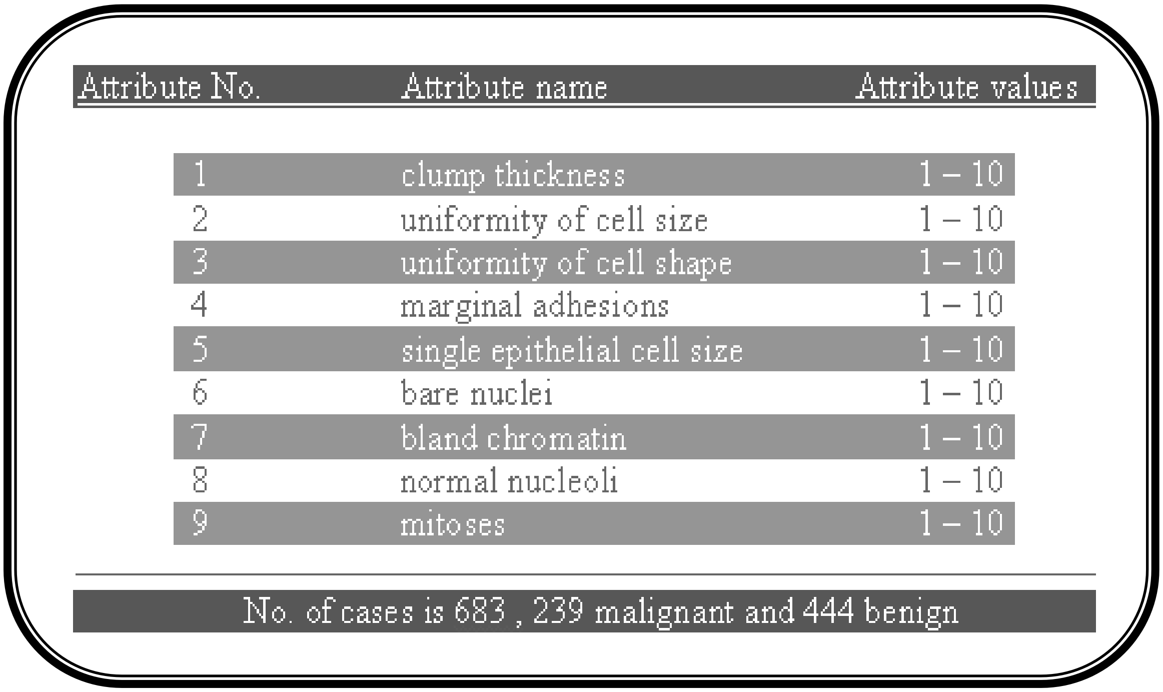


Figure 1: An example for clustering of randomly generated dataset

4. WBCD description

A database consisting of 699 cases is prepared by Dr. William H. Wolberg in the University of Wisconsin Hospital [Htet, 2015]. The data in the WBCD is taken from real cases of patients who underwent FNA biopsies [Ashutosh Patri ,2014].

The database is made up of nine symptoms. The value of each symptom ranges from 1 to 10 with class status and who is either benign (2) or malignant (4). There are 16 cases containing missing values, and after the filtering these cases by deleting them, the remaining is 683 cases, including 444 benign cases and 239 malignant instances [Hussein, 2013]. The information of WBCD is summarized in figure 2.



Attribute No.	Attribute name	Attribute values
1	clump thickness	1 – 10
2	uniformity of cell size	1 – 10
3	uniformity of cell shape	1 – 10
4	marginal adhesions	1 – 10
5	single epithelial cell size	1 – 10
6	bare nuclei	1 – 10
7	bland chromatin	1 – 10
8	normal nucleoli	1 – 10
9	mitoses	1 – 10

No. of cases is 683 , 239 malignant and 444 benign

Figure 2 : WBCD description

5. K-means for WBCD problem

In this section, a description of K-means parameter for breast cancer is presented. Figure 3 shows the steps of the K-means clustering for WBCD:

- The No. of clusters: for this problem, we need two clusters since the dataset has two class, i.e. cluster for benign cases and cluster for malignant instances.
- Choosing the center of the cluster: this is done by selecting two instances randomly, provided that the classes for each are different to become a magnet for similar cases.
- The class of the cluster: the class of the initial centers represent the class of the cluster.
- Distance measure : Euclidean distance is used as a measure to detect similarity between the case and the cluster's center, i.e. for each case, Euclidean distance is applied with all the clusters centers and then allocate the case to the cluster with the minimum distance.
- Updating the centers: When all cases in the database are grouped into clusters, the algorithm announces the end of the cycle and then update each center by averaging the point belonged to that cluster .
- Stop criterion: the proposed algorithm stops when all instances settle in clusters permanently, in other words, the instances are not transmitted from cluster to another.

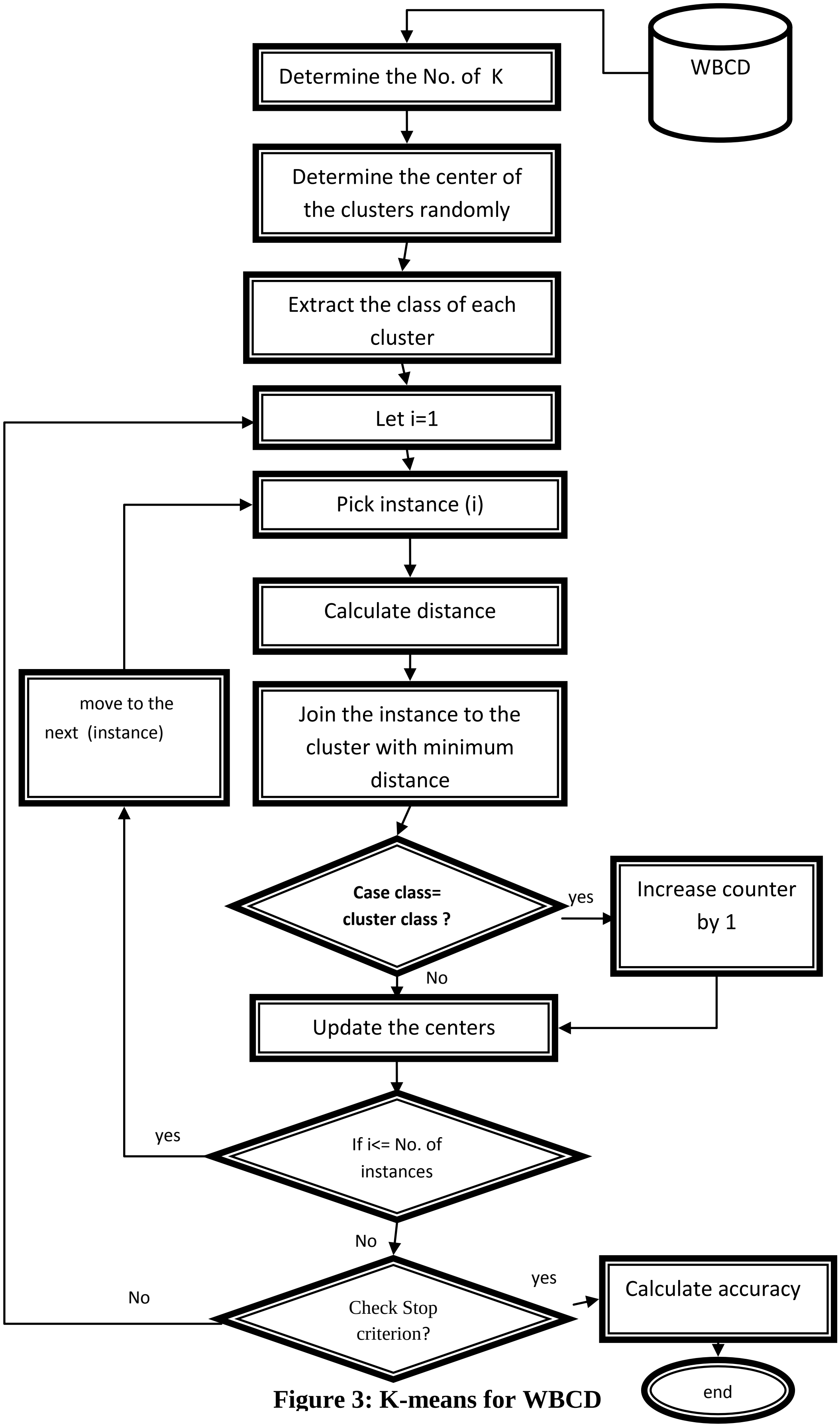


Figure 3: K-means for WBCD

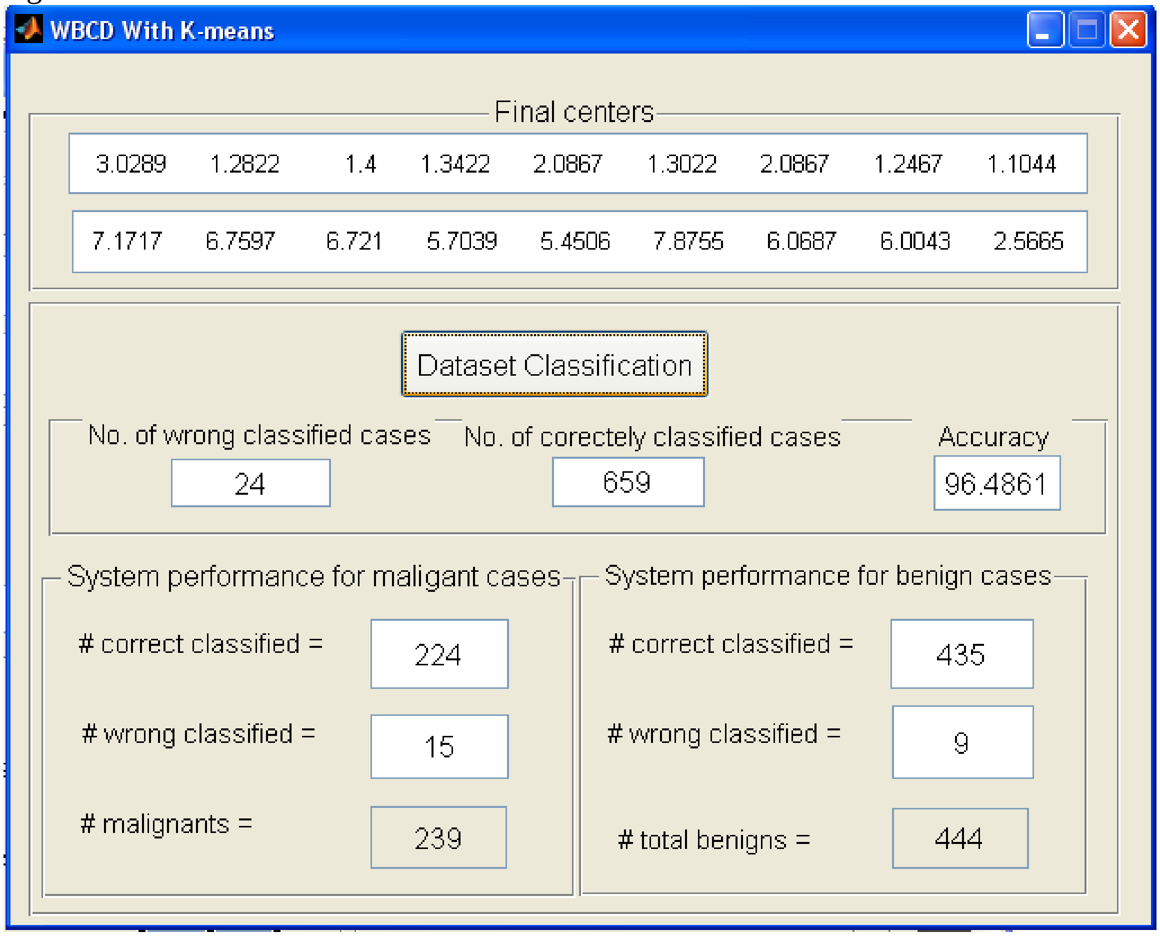
- Calculating classification ratio: This is done based on the following equation:

$$Performance = \frac{No.of\ correctley\ classified\ instances}{683} \times 100\ \%$$

After the allocation of an instance to a certain cluster, a comparison between the class of instance in database and the class of the cluster which belonged to. If they are identical a counter is incremented by one to refer to the success of the method in classifying this case. At the end, the algorithm calculates the number of cases classified correctly and divides this number by the total cases (683).

6. Results

Matlab R2011a is used to perform the proposed idea. The implementation reveals the success of the system in classifying the database by up to 96.4861% as shown in Figure 4:



Final centers

```
3.0289   1.2822   1.4     1.3422   2.0867   1.3022   2.0867   1.2467   1.1044
7.1717   6.7597   6.721   5.7039   5.4506   7.8755   6.0687   6.0043   2.5665
```

No. of wrong classified cases: 24
No. of corectely classified cases: 659
Accuracy: 96.4861

System performance for maligant cases
# correct classified = 224
# wrong classified = 15
# malignants = 239

System performance for benign cases
# correct classified = 435
# wrong classified = 9
# total benigns = 444

Figure 4: The result of the proposed system

The figure presents the final center for benign cluster which is:

[3.0289 1.2822 1.4 1.3422 2.0867 1.3022 2.0867 1.2467 1.1044]

And the final center for malignant cluster which is :

[7.1717 6.7597 6.721 5.7039 5.4506 7.8755 6.0687 6.0043 2.5665].

The following table shows some important details about the performance of the proposed system:

Table 1: performance measures for proposed system

performance measures	Number of instances	Total instances
correctley classified instances	659	683
wrong classified instances	24	
correctley classified benign *instances*	435	444
wrong classified benign *instances*	9	
correctley classified malignant *instances*	224	239
wrong classified malignant *instances*	15	

7. Conclusion and future work

This work sheds light on the use of K-means technology to classify a global database dedicated for cancer research that is WBCD. The method is demonstrated its ability to classify the database successfully and it is managed to get a high performance despite its simplicity. For future work, it is possible to strengthen the performance of method through the use of optimization methods that elevate the performance of the system, and the initial idea is to use genetic algorithm to do one of the following operations:

- Reducing the number of features and get the best and most influential nomination.
- Selection of the initial centers of the clusters.

References

Ashutosh Patri, Abhijit Nayak, 2014. " High accuracy back-retreat diffusion-fuzzy clustering of breast cancer data for the detection of malignancy".

Bhagwati Charan Patel, G.R.Sinha,2010. " An adaptive k-means clustering algorithm for breast image segmentation", international journal of computer applications, vol. 10, no.4, p.p. 0975 – 8887.

Dharmendra K Roy , Lokesh K Sharma, 2014."Genetic k-means clustering algorithm for mixed numeric and categorical data sets ", international journal of artificial intelligence and applications,vol.1,no.2.

Htet Thazin Tike Thein and Khin Mo Mo Tun, 2015." An approach for breast cancer diagnosis classification using neural network", advanced computing: an international journal (acij), vol.6, no.1.

Hussein A.Lafta , Noor K. Ayoob,2013. "Breast cancer diagnosis using genetic fuzzy rule based system", journal of Babylon university/pure and applied sciences, vol.21, no.4.

Indira Muhic, 2013. " Fuzzy Analysis of Breast Cancer Disease using Fuzzy c-means and Pattern Recognition ", Southeast Europe Journal Of Soft Computing, Vol. 2, No. 1.

Karabatak M. 2009. "An expert system for detection of breast cancer based on associative rules and neural network", expert system with Applications,Vol. 36,pp. 3465-3469.

Mahua Nandy, 2013."An Analytical Study of Supervised and Unsupervised Classification Methods for Breast Cancer Diagnosis " , 2nd International conference on Computing Communication and Sensor Network.

Nikhil Chaturvedi , Er. Anand Rajavat, 2013. "An improvement in k-mean clustering algorithm using better time and accuracy ", International journal of programming languages and applications (ijpla), vol.3, no.4.

Nikita Ghiya, and etal, 2015." Forecasting diseases by classification and clustering techniques", international journal of innovative research in science,engineering and technology, vol. 4, no.1.

Sridevi, A.murugan, 2014. "An intelligent classifier for breast cancer diagnosis based on k-means clustering and rough set ", international journal of computer applications, vol. 85 , no. 11,p.p , 0975 – 8887.

Thakare and S. B. Bagal, 2015." Performance evaluation of k-means clustering algorithm with various distance metricsy", international journal of computer applications, vol. 110, no. 11,p.p.0975 – 8887.

Wanli Xiang and etal, 2015." A dynamic shuffled differential evolution algorithm for data clustering".