# Data Mining based Prediction of Medical data Using K-means algorithm

***Dr. Bushra M. Hussan***

***Computer Science Department - College of Science - Basrah University***

**ABSTRUCT**

Data mining is one of the knowledge discovery steps in database, in which modelling techniques are applied. In this paper, K-means method is applied for dealing with medical database for clustering. To increase the efficiency of mining process, some pre-processing need to be done to the data. Experimental results showed the good accuracy when applied to the adjust data.

**Keywords**: Data mining, Knowledge Discovery, K-means, Clustering.

## INTRODUCTION

The data mining technique has become an established method for improving statistical tools to predict future trends [Plamena].

It is a process of semi-automatically analyzing large databases to find pattern that are valid, novel, useful and understandable [Runumi]

The aim of this study is to test the K-means algorithm which is one of data mining techniques at medical data bases.

Cluster analysis is one of the major data analysis which helps to identify the natural grouping in a set of data item[K.A].

Clustering is operation or process of partitioning a given set of objects into disjoint clusters[K.A].

There are several motivation for clustering [Cambridge]:

**First,** a good clustering has predictive power. Since cluster labels are meaningful which lead to more efficient description of given data, and will help to choose better actions

**Second,** clusters can be a useful aid to communication because they allow lossy compression. In lossy compression, the aim is to convey in as few bits as possible a reasonable reproduction of a picture; one way

to do this is to divide the image into N small

in an alphabet of K image-templates, then we send a close fit to the image by sending the list of labels of the matching template.

**Third,** failures of the cluster may highlight interesting objects that deserve special attention.

**Forth**, clustering algorithms may serve as models of learning process in neural systems.

## MEDICAL DATA CHARACTERISTIC

Raw medical data are voluminous and heterogeneous and may be collected from various images, interviews with the patient and the physician's observations and interpretations[Krzysztof]. The medical data are characterized by their incompleteness (missing parameter value), incorrectness (noise in data Nearly all diagnoses and treatment in ), sparseness (few and/or non-representable patient record are available) and patches, and find a close match to each patch inexactness (inappropriate selection of parameters for a given task) . One should distinguish between a test and a diagnosis , a test is one of many values used to characterize the medical condition of patient; a diagnosis is the synthesis of many tests and observation, that describes a pathophysiologic process in that patient , both tests and diagnoses are subject to sensitivity / specificity analysis[Sarojini].

## DATASET[Sarojini]

The experiments were performed on the Pima Indian diabetes dataset from the UCI (University of California at Irvine), which consist of 768 complete instances described by 8 feature(**Glucose tolerance test, Diastolic blood pressure, Triceps skin fold thickness, 2Hour serum insulin, Body mass index, Diabetes pedigree function, Age, 1 is tested positive**).

**Table 1.represent part of the original dataset**

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 140 | 94 | 0 | 0 | 32.7 | 0.734 | 45 | 2 |
| 108 | 80 | 0 | 0 | 27 | 0.259 | 52 | 2 |
| 128 | 48 | 45 | 194 | 40.5 | 0.613 | 24 | 2 |
| 130 | 82 | 0 | 0 | 39.1 | 0.956 | 37 | 2 |
| 121 | 66 | 30 | 165 | 34.3 | 0.203 | 33 | 2 |
| 109 | 64 | 44 | 99 | 34.8 | 0.905 | 26 | 2 |
| 145 | 80 | 46 | 130 | 37.9 | 0.637 | 40 | 2 |
| 123 | 62 | 0 | 0 | 32 | 0.226 | 35 | 2 |
| 151 | 90 | 46 | 0 | 42.1 | 0.371 | 21 | 2 |
| 130 | 70 | 0 | 0 | 34.2 | 0.652 | 45 | 2 |
| 146 | 0 | 0 | 0 | 27.5 | 0.24 | 28 | 2 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 173 | 70 | 14 | 168 | 29.7 | 0.361 | 33 | 2 |
| 111 | 72 | 47 | 207 | 37.1 | 1.39 | 56 | 2 |
| 113 | 76 | 0 | 0 | 33.3 | 0.278 | 23 | 2 |
| 135 | 68 | 42 | 250 | 42.3 | 0.365 | 24 | 2 |
| 137 | 108 | 0 | 0 | 48.8 | 0.227 | 37 | 2 |
| 121 | 52 | 0 | 0 | 36 | 0.127 | 25 | 2 |
| 151 | 78 | 32 | 210 | 42.9 | 0.516 | 36 | 2 |
| 122 | 56 | 0 | 0 | 33.3 | 1.114 | 33 | 2 |
| 128 | 68 | 19 | 180 | 30.5 | 1.391 | 25 | 2 |
| 197 | 74 | 0 | 0 | 25.9 | 1.191 | 39 | 2 |
| 172 | 68 | 49 | 579 | 42.4 | 0.702 | 28 | 2 |
| 128 | 78 | 37 | 182 | 43.3 | 1.224 | 31 | 2 |

**THE K-MEANS ALGORITHM [K.A, Xindong]**

The k-means algorithm is a simple iterative method to partition the given dataset into user-specified number of clusters, k.

This function depends on fuzzy logic in its work it assumes many centers, then finds the smallest distance from the points to these centers, then rearrange these points as clusters. The distance of each cluster from fixed center is less than the distance from other centers.

The algorithm operates on a set of d-dimensional vectors. The algorithm is initialized by picking k points as the initial k cluster representative. Techniques for selecting these initial seeds include sampling at random from the dataset, setting them as the solution of clustering a small subset of the data or perturbing the global mean of the data k times. Then the algorithm iterates between two steps till convergence:

Step 1:Data assignment: Each data point is assigned to its closest centroid, with ties broken arbitrarily. This result is partitioning of the data.

Step 2:Relocation of "means". Each cluster representative is relocated to the center (mean) of all data points to it. If the data come with a probability measure (weights), then the relocation is to the expectations (weighed mean) of the data partitions.

Once issue to resolve is how to quantify "closest" in the assignment steps. The default measure of closeness is the Euclidean distance, in which case one can readily show that the non-negative cost function:

$$\sum \arg\min(xi - cj) \qquad (1)$$

Will decrease whenever there is a change in the assignment or relocation steps, and hence convergence is guaranteed in a finite number of iteration.

**CHOOSING THE NUMBER OF CLUSTERING[8]**

One of the main disadvantage to k-means is the fact that specified the number of clusters as an input to the algorithm, The algorithm is not capable of determining the appropriate number of clusters and depends upon the user to identify it. It's a good idea to experiment with different values of k to identify the value that best suits the data.

**THE ACCURACY MEASURMENT[Hnin]**

The accuracy of the test compares how close a new test value to a value predicted. An accuracy test is defined as:

$$\text{Accuracy} = \frac{TP}{total} 100\%$$

where TP stands for true positive and indicates the number of correctly recognized test example and total is the total number of test example.

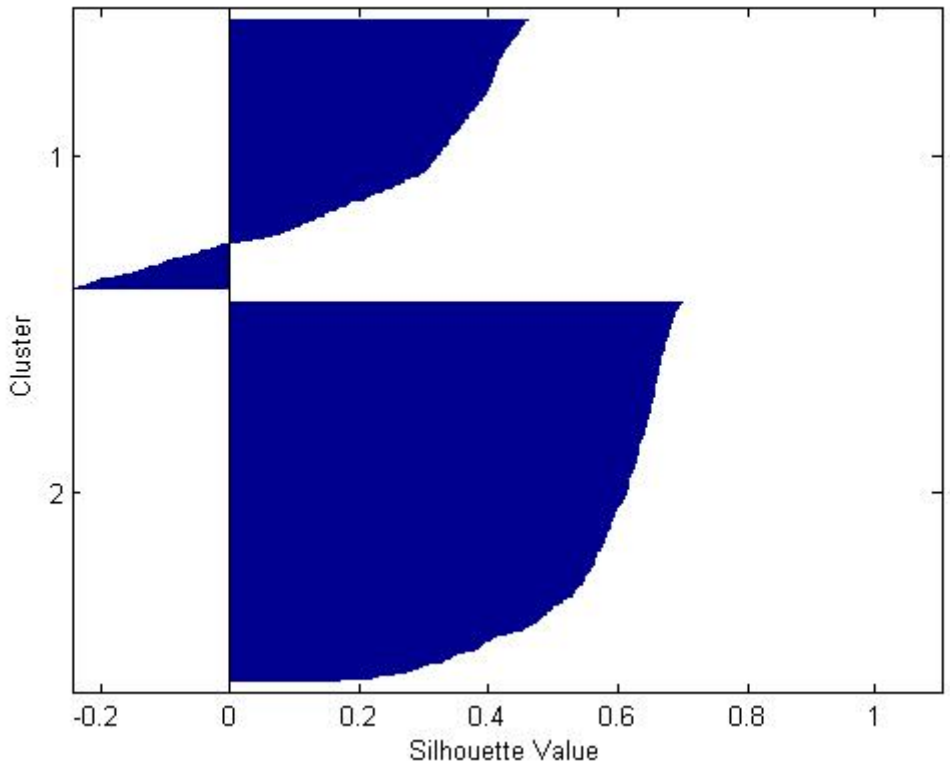


**Figure 1.The original data in clusters**

**Table 2. represent the number of the cluster which each instances belongs to. Preprocessing[Edgar]**

| **Idx** |
|---|
| 2 |
| 2 |
| 1 |
| 2 |
| 1 |
| 1 |
| 1 |
| 2 |
| 2 |
| 2 |
| 2 |
| 1 |
| 1 |
| 2 |
| 1 |
| 2 |
| 2 |
| 1 |
| 2 |
| 1 |
| 2 |
| 1 |
| 1 |

The Diabetes Pima database which used in this research contain missing data which is a common problem in statistical analysis. Rates of less than 5% is manageable, 5-15% require sophisticated methods to handle, and more than 15% may severely impact any kind of interpretation, several methods have been proposed to treat missing data :

a) Case Deletion:

It is a default method in many programs, This method consists of discarding all instance(cases) with missing values for at least one feature of the database.

b )Mean Imputation:

This is one of the most frequently used methods, It consist of replacing the missing data for a given feature (attribute) by the mean of all known values of that attribute.

Let us consider that the value x of the k-th class, C , is missing then it will replaced by

$$x = \sum x / n \qquad (2)$$

Where n represents the number of non-missing values in the j-th feature of the k-th class.

c) Median Imputation:

In this method the missing data is replaced by the median of all known values of that attribute in the class, the method also recommended choice when the distribution of the values of a given feature is skewed. Let us consider that the value x of the k-th class, C , is missing, it will be replaced by

$$x = \text{median} \{x\} \qquad (3)$$

d) KNN Imputation:

The algorithm of this method is as follows:

1.Divide the data set into two parts, Let D be the set containing the instance in which at least one of the features is missing, The remaining instances will complete feature information from a set called D .

2.For each vector x in D :

-Divide the instance vector into observed and missing parts as x={x ; x }.

-Calculate the distance between the x and all the instance vectors from the set D . Use only those features in the instance vectors from the complete set D which are observed in the vector x.

-Use the K closest instance vectors(K-nearest neighbors) and perform a majority voting estimate of the missing values for categorical attributes.

We apply the last one to modify the data set and remove the missing values, The result is as shown in Table 3, and the data distributed in the clusters as in Figure 2.

**Table 3.represent the same part of the dataset after the preprocessing**

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 140 | 94 | 44 | 168 | 32.7 | 0.734 | 45 | 2 |
| 108 | 80 | 44 | 168 | 27 | 0.259 | 52 | 2 |
| 128 | 48 | 45 | 194 | 40.5 | 0.613 | 24 | 2 |
| 130 | 82 | 44 | 168 | 39.1 | 0.956 | 37 | 2 |
| 121 | 66 | 30 | 165 | 34.3 | 0.203 | 33 | 2 |
| 109 | 64 | 44 | 99 | 34.8 | 0.905 | 26 | 2 |
| 145 | 80 | 46 | 130 | 37.9 | 0.637 | 40 | 2 |
| 123 | 62 | 44 | 168 | 32 | 0.226 | 35 | 2 |
| 151 | 90 | 46 | 168 | 42.1 | 0.371 | 21 | 2 |
| 130 | 70 | 44 | 168 | 34.2 | 0.652 | 45 | 2 |
| 146 | 70 | 44 | 168 | 27.5 | 0.24 | 28 | 2 |
| 173 | 70 | 14 | 168 | 29.7 | 0.361 | 33 | 2 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 111 | 72 | 47 | 207 | 37.1 | 1.39 | 56 | 2 |
| 113 | 76 | 44 | 168 | 33.3 | 0.278 | 23 | 2 |
| 135 | 68 | 42 | 250 | 42.3 | 0.365 | 24 | 2 |
| 137 | 108 | 44 | 168 | 48.8 | 0.227 | 37 | 2 |
| 121 | 52 | 44 | 168 | 36 | 0.127 | 25 | 2 |
| 151 | 78 | 32 | 210 | 42.9 | 0.516 | 36 | 2 |
| 122 | 56 | 44 | 168 | 33.3 | 1.114 | 33 | 2 |
| 128 | 68 | 19 | 180 | 30.5 | 1.391 | 25 | 2 |
| 197 | 74 | 44 | 168 | 25.9 | 1.191 | 39 | 2 |
| 172 | 68 | 49 | 579 | 42.4 | 0.702 | 28 | 2 |
| 128 | 78 | 37 | 182 | 43.3 | 1.224 | 31 | 2 |

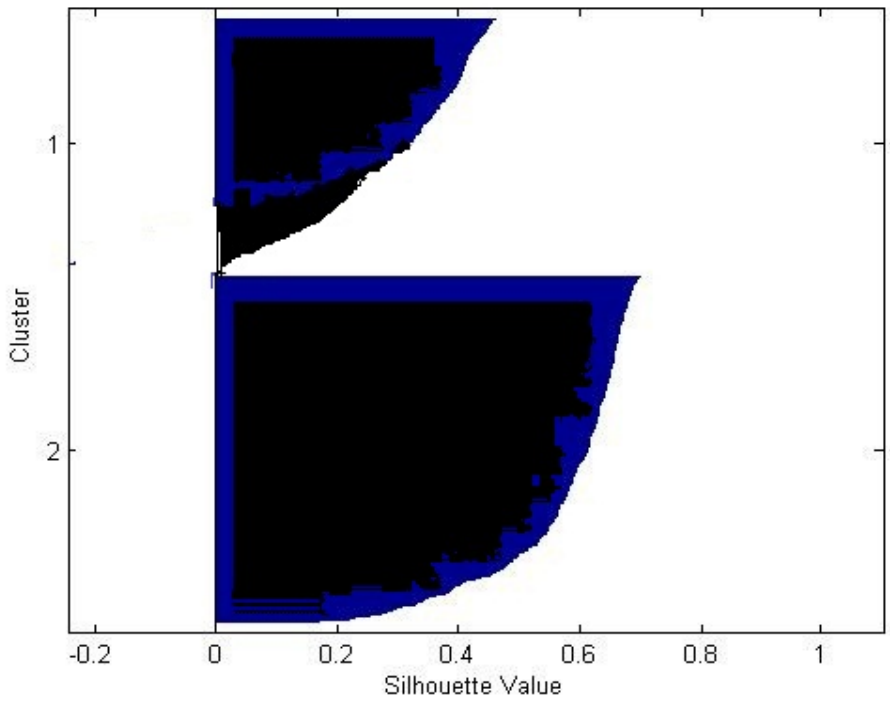


**Figure 2. The clusters after the preprocessing of the data**

**Table 4.represent the new dataset for testing**

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 142 | 94 | 44 | 168 | 32.7 | 0.734 | 45 | 2 |
| 108 | 81 | 44 | 168 | 28 | 0.250 | 52 | 2 |
| 129 | 48 | 44 | 190 | 40.3 | 0.613 | 24 | 2 |
| 130 | 81 | 44 | 168 | 39.3 | 0.954 | 37 | 2 |
| 121 | 66 | 31 | 165 | 34.3 | 0.203 | 33 | 2 |
| 119 | 64 | 44 | 100 | 34.1 | 0.905 | 26 | 2 |
| 145 | 81 | 45 | 130 | 37.5 | 0.631 | 40 | 2 |
| 123 | 62 | 44 | 168 | 32 | 0.226 | 35 | 2 |
| 151 | 90 | 46 | 168 | 42.2 | 0.371 | 21 | 2 |
| 131 | 70 | 44 | 168 | 34.2 | 0.652 | 45 | 2 |
| 146 | 80 | 44 | 168 | 27.5 | 0.243 | 28 | 2 |
| 173 | 70 | 14 | 168 | 29.7 | 0.361 | 33 | 2 |
| 111 | 77 | 48 | 217 | 36.3 | 1.391 | 56 | 2 |
| 114 | 77 | 44 | 168 | 33.3 | 0.278 | 23 | 2 |
| 135 | 68 | 40 | 240 | 42.3 | 0.365 | 24 | 2 |
| 137 | 118 | 44 | 168 | 47.8 | 0.227 | 37 | 2 |
| 124 | 52 | 44 | 168 | 34 | 0.128 | 25 | 2 |
| 152 | 79 | 32 | 210 | 42.4 | 0.516 | 36 | 2 |
| 130 | 56 | 44 | 168 | 31.3 | 1.114 | 33 | 2 |
| 126 | 67 | 19 | 180 | 30.5 | 1.392 | 25 | 2 |
| 197 | 74 | 44 | 168 | 25.4 | 1.191 | 39 | 2 |
| 172 | 68 | 48 | 580 | 42.4 | 0.702 | 28 | 2 |
| 128 | 78 | 37 | 182 | 43.3 | 1.224 | 31 | 2 |

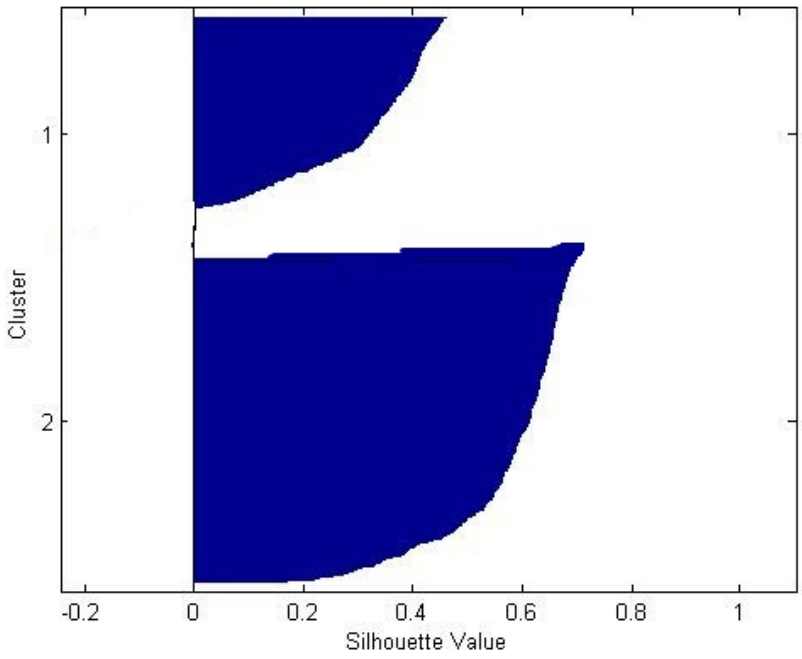


**Figure 3.The clusters represent the testing data.**

**Table 5. represent the number of the cluster which each instances of testing data belongs to.**

| Idx |
|---|
| 2 |
| 1 |
| 1 |
| 2 |
| 2 |
| 1 |
| 1 |
| 2 |
| 1 |
| 2 |
| 1 |
| 1 |
| 1 |
| 1 |
| 1 |

2
2
1
2
1
2
1
2

## CONCLUSIONS AND DISCUSSION

The K-mean algorithm is widely used for clustering large sets of data, We try many values for k, lastly we find that two clusters for the data is the best one.

Part of the original Pima dataset is shown in Table 1.

Table 2 is shown the number of which cluster each instance of the original data belongs.

The Pima dataset is preprocessed successfully by supplying missing values as shown in table 3 using the KNN mutation , then clustered using K-means with k value equal 2.

Figure 1 and Figure 2 present the clusters include all the data (768 instance) before and after the preprocessing.

The first result of algorithm execution on the original data showed that we get accuracy 81% which is not good rate, So the data was improved by the preprocessing process and then apply the algorithm again, it gave us accuracy 94%.

When we want to examine the algorithm, we apply it with new instances(700 records), Table 4 represent part of it. and notice the number (idx) of the cluster as in Table 5, in which the instances lay, and we get the accuracy 97%.

## REFERENCES

Edgar Acuna and Caroline Rodriguez, " The treatment of missing values and its effect in the classifier accuracy", Department of mathematics, University of Puerto Rico Mayaguez, , PR00680edgar@cs.uprm.edu.

Hnin Wint Khaing "Data Mining based Fragmentation and Prediction of Medical Data", University of computer studies, Mandalay, snow.hwk@gmail.com, 2011.

K A Abdul Nazeer, S D Madhu Kumar, "Enhancing the K-means clustering algorithm by using O(nlog) heuristic method for finding better initial centroids", computer society IEEE, nazeer@nitc.ac.in, madhu@nitc.ac.in, 2011.

Krzysztof J. Cios and G. William Moore,"Uniqueness of Medical Data Mining ",apper in Artificial Intelligence in Medicine journal, 2002.

Plamena Andreeva, Maya Dimitrova , " Data Mining Learning Models and Algorithms for Medical Applications", plamena@icsr.bas.bg, dimitrova@icsr.bas.bg .
Computer Vision Center, Autonomous University Barcelona, Spain, peta@cvc.uab.es ,2003.

Runumi Devi and Vineeta Khemchandani, " Application of Data Mining Techniques For Diabetic Dataset", Proceeding of the 4th National Confrence;INDIACom-2010.

Sarojini Balakrishnan," Feature Selection Using FCBF in Type 2 Diabetes Databases",International Conference on IT Celebrate S. balakrishnan.sarojini@gmail.com, 2009.

Xindong We, Vipin Kumar, " Top 10 Algorithms in Data Mining",2007.

**Cambridge University, http://www.cambridge.org, 2003.**