# Comparison of Performance Between Back Propagation and K-means on Medical Datasets

**Asraa Abaullah Hussen**
*Science Collage for Women / Computer Science/University of Babylon*
esraa_zd@yahoo.com

## Abstract:

In recent decades, and to this day computer technology has been used in applications and various fields including the medical field, which prompted many researchers to employ this technique in the design of decision support systems using many of the algorithms and methods for this purpose. In this paper, k-means and back propagation are proposed to classify medical datasets and then compare the performance of these methods, practical experiments show back propagation has best results than k-means.

**Keywords** : Medical datasets, k-means, back propagation, diseases.

## الخلاصة:

في العقود الأخيرة، وحتى يومنا هذا تكنولوجيا الحاسوب استخدمت في تطبيقات ومجالات مختلفة ومن ضمنها المجال الطبي، الذي دفع العديد من الباحثين إلى توظيف هذه التقنية ببناء أنظمة دعم القرار من خلال تطبيق العديد من الخوارزميات والطرق لهذا الغرض. في هذا البحث الشبكات ألعصبيه وK-means أقترحت لتصنيف القواعد الطبية ومن ثم مقارنة أداء هذه الطرق، التجارب العملية أظهرت الشبكات العصبية تمتلك أداء أفضل من k-means .

**الكلمات المفتاحية:-** قواعد طبية، معدلات معامل(K-means)، التغذية الرجعية، أمراض.

## 1 Introduction:

Data mining seeks a solution for real world health problems in the diagnosis and handling diseases. Several data mining techniques were used by Researchers in the medical field such as decision tree, k-means, fuzzy c-means, k-nn and neural network(Das, 2009).

To decrease cost and human effects K.Rajalakshmi, Dr.S.S.Dhenakaran and N.Roobini proposed a prediction system. This system includes analyze different sickness prediction by using k-means clustering algorithm, for this purpose three medical dataset were used (Heart Disease, Diabetics, Liver disease and Cancer)(Rajalakshmi., 2015).

Nitu M. and Ashish B. implemented a classifier system containing k-means and back propagation algorithms. They conducted a study by applying two methods on the staffing data of an organization to analyze the performance of each method, at end the result of study show that back propagation is better than k-means(Nitu, 2012).

Heart disease is the most common factor for death in India, in order to reduce the risk of this disease, the trend has been to design decision support systems to help doctors diagnose heart disease process with less features. From this standpoint, the researchers Priti , M. A. jabbar and B.L Deekshatulua Chandra went to design the system of diagnosis heart disease based on k-nn and genetic algorithm. They use k-nn as a classifier method and genetic algorithm for reduce the features. The result of system proves high accuracy(Priti, 2013).

## 2 Classifier methods:

There are several algorithms and methods in the field of diagnosis and classification for Medical datasets. In this paper, we have been focus on k-means and back propagation neural.

## 2.1 K-means algorithm:

Clustering is an operation of splitting a dataset into sets such that the individuals of each set are similar with each other as much as possible and different with individuals of other sets. In cluster analysis, there is no previous knowledge about the elements belonging to groups. The elements are grouped through data analysis(Mouslem, 2011).

K-means algorithm is a way of partitioning methods developed in 1967 by James Macqueen, and is widely used because it is Characterized by the ease and simplicity(Swarndeep, 2016). The algorithm can be explained over a sets of steps as follow:

Algorithm : K-means clustering:

**Input:** datasets containing number of objects.

**Output:** divided dataset into k of groups.

**Step1:** Starting determine the number of groups(k).

**Step2:** Pick from dataset k of objects represent centers of each group randomly.

**Step3:** Compute the distance between the object and center of each group, based on the result of distance object allocated to a group with small distance. Distance measure that are used is Manhattan Distance which the equation as following:

$$\text{Manhattan Distance} = \sum_{f=1}^{k} |\mathbf{af} - \mathbf{zf}|$$

**Step4:** Update the center of each group by compute the mean of values.

**Step5**: Stop when there is no change, else go to step 3.

## 2.2 Back Propagation Neural Networks:

In 1969, Bryson and Ho was invented Back propagation (generalized delta rule) as a way for learning in multi-layer network.

The back propagation algorithm used to calculate the major repairs after the random selection for the weights of the network.

Description of Training BP Net:( Abbas, 2015)

Feed forward Phase

1. At begging, creating a little random values for weights.

2. As long as the stop state hasn't been achieved, work the following for every training pair (input/output):

- Each value from the input unit passes to all hidden units
- sums input signals for each hidden unit and calculate its output signal by applying the activation function.
- send the value of each hidden unit to the output units
- sums input signals for each output unit and calculate its output signal by applying the activation function.

Back propagation Phase

3. For each output calculate the fault and correct its own( weight, bias) then transmit it to layer below

4. Collect the inputs of each hidden unit from above and multiplies with the derivative of its activation function; also computes its own( weight, bias) correction.

5. For each output unit, updates the weights and bias.

6. For each hidden unit, updates the weights and bias ( Youssef , 2012).

## 3 The Results of suggested work:

This section deals with clarifying the results of the proposed work, which consists of the application of k-means and back propagation algorithms on medical datasets to analysis the difference in performance. For this purpose, three medical datasets were used for experimental (Breast cancer, Heart disease and Diabetes data sets) are taken from the UCI machine learning repository as shown in Table (1). Table(2) and Table(3) show the results of applying two methods.

Table(1): Database Information.

| Dataset | No. of instance | No. of features | Class | No. of patient and not patient |
|---|---|---|---|---|
| **Breast disease** | 684 | 9 | Integer valued 2 (benign) and 4 (malignant) | 444 benign<br>239 malignant |
| **Heart disease** | 270 | 13 | 0 not patient<br>1 patient | 150 not patient<br>120 patient |
| **Diabetes disease** | 85 | 26 | 1 not patient<br>2 patient | 38 not patient<br>47 patient |

Table (2): Result of back propagation.

| Disease | No. of features/No. of input cells | Performance of (80%train-20%test) | Performance of (70%train - 30%test) | Performance of (50%train - 50%test) |
|---|---|---|---|---|
| **Heart** | 13 | 83.3333% | 87.6543% | 83.7037 |
| **Breast cancer** | 9 | 98.5401% | 99.0244% | 96.7742% |
| **Diabetes** | 26 | 100% | 96.153855% | 88.0952% |

Table (3): Result of k-means.

| Disease | Performance when applying k-means |
|---|---|
| **Heart** | 67% |
| **Breast cancer** | 95% |
| **Diabetes** | 74% |

## 4 Conclusion:

Set of experiments have conducted on the medical datasets as described in table(2) and table(3) to view the performance of back propagation compared with K-means. In back propagation each dataset has been divided into train and test differ in size as mentioned in the table(2) , the performance of this method in range of (83 - 100).The back propagation shares the same architecture one cell for output, hidden layer contain 5 cells and number of input cells based on features of disease means equal to number of features. K-means display a good results but not the level of the back propagation.

## 5 Reference:

Abbas H. Issa, 2015, "Artificial Neural Networks Based Fingerprint Authentication", Eng. &Tech.Journal, Vol.33,Part (A), No.5.

Das, R., I. Turkoglu, and A. Sengur, 2009,"Effective diagnosis of heart disease through neural networks ensembles. Expert Systems with Applications", Elsevier, 36: p. 7675–7680.

Rajalakshmi, Dr. S. S. Dhenakaran and N. Roobini, 2015," Comparative Analysis of K-Means Algorithm in Disease Prediction", International Journal of Science, Engineering and Technology Research (IJSETR), Volume 4, Issue 7, July.

Priti, M. A. jabbar and B.L Deekshatulua , 2013, " Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm", International Conference on Computational Intelligence: Modeling Techniques and Applications (CIMTA), Procedia Technology 10 85 – 94.

Moslem M. K., and Abbas H. H. A., 2011, " Applying New Method For Computing Initial Centers Of K-means Clustering With Color Image Segmentation", J.Thi-Qar Sci., Vol. 3, No. 1.

Nitu M. and Ashish B., 2012," Comparison of K-means and Back propagation Data Mining Algorithm", International Journal of Computer Technology and Electronics Engineering (IJCTEE) Volume 2, Issue 2, April 15.

Swarndeep P. and S. Saket J., 2016,"Implementation of Extended K-Medoids Algorithm to Increase Efficiency and Scalability using Large Datasets ", International Journal of Computer Applications (0975 – 8887) Volume 146 – No.5, July.

Youssef B., 2012,Neural Network Model for Path-Planning Of Robotic Rover Systems, International Journal of Science and Technology (IJST), E-ISSN: 2224-3577, Vol. 2, No. 2, February.