International Journal of Electrical and Computer Engineering (IJECE)
Vol. 8, No. 2, April 2018, pp. 1256~1261
ISSN: 2088-8708, DOI: 10.11591/ijece.v8i2.pp1256-1261

Improve The Performance of K-means by using Genetic Algorithm for Classification Heart Attack

Asraa Abdullah Hussein
Department of Computer Science, Science Collage for Women, University of Babylon, Iraq

Article Info

Article history:

Received Nov 26, 2016
Revised Jan 2, 2018
Accepted Dec 13, 2018

Keyword:

Features
Genetic algorithm
Heart disease
K-means

ABSTRACT

In this research the k-means method was used for classification purposes after it was improved using genetic algorithms. An automated classification system for heart attack was implemented based on the intelligent recruitment of computer capabilities at the same time characterized by high performance based on (270) real cases stored within a globally database known (Statlog). The proposed system aims to support the efforts of staff in medical felid to reduce the diagnostic errors committed by doctors who do not have sufficient experience or because of the fatigue that the doctor suffers as a result of work pressure. The proposed system goes through two stages: in the first-stage genetic algorithm is used to select important features that have a strong influence in the classification process. These features forms the inputs to the K-means method in the second-stage which uses the selected features to divide the database into two groups one of them contain cases infected with the disease while the other group contains the correct cases depending on the distance Euclidean. The comparison of performance for the method (K-means) before and after addition genetic algorithm shows that the accuracy of the classification improves remarkably where the accuracy of classification was raised from (68..1481) in the case of use (k- means only) to (84.741) when improved the method by using genetic algorithm.

Corresponding Author:

Asraa Abdullah Hussein,
Department of Computer Science,
Science Collage for Women,
University of Babylon, Iraq.
Email: esraa_zd@yahoo.com

1. INTRODUCTION

The tremendous progress that has accompanied computer science and the success it has achieved in various applications has made it more than just a computing machine and this has been a powerful motivation for scientists to develop and invent several technologies that try to exploit the capabilities of the computer to accomplish useful functions and find solutions to many problems to facilitate the joints of human life and reduce the problems that may be faced so many techniques have emerged including: (expert systems, networks and classification algorithms of various types) [3] .

Classification of diseases is a distinctive goal of artificial intelligence research that has tried to support the medical field and provide specialists of doctors, centers and hospitals with diagnostic systems that help to improve the accuracy of decision made on a situation and reduce errors that may be made in the diagnosis because of lack of experience or pressure stress which leads to problems in the accuracy of the diagnosis for specialist and also provides detailed medical data about the test in record time [6-8].

The heart attack is one of the dangers diseases that threaten human life where The World Health Organization (WHO) reports that 12 million people die each year from heart disease [1]. Because the severity of disease many computer specialists presented on many years a lot of research aimed to supporting medical institutions and their staff with systems to diagnose this disease and research is still ongoing in the field [5].

Researchers rely on a global database known as (Statlog). This database used in research that work on classification heart attack to measure the strength of the method proposed by the research. It can be obtained from the data warehouse (UCI) allocated each row in this database for each patient. The total number of cases (patients) in the database are (270) case and each person stored 13 information (property): (age), (sex), (chest pain type), (blood pressure), (cholesterol), (blood sugar), (electrocardiographic results), (maximum heart rate) and other properties. The property 14 is represent the final diagnosis: the value of this property is (1) to indicate for infected person while the healthy person referred by making the value of property 14 equal to (0). Table 1 summarizes the important and most recent research that classified this disease by categorizing the database (Statlog) the data set in the table sorted by year of publication.

Table 1. Summary of previous relevant research

accuracy	Methods used	Researcher Name and Year of Publication
86%	Navie bayes and laplsing smothing	[Vincy Cherian 2017]
83.5165%	Bayes Net	[Bharti Dansena 2017]
75.8889 %	Chaotic Particle Swarm Optimization	Zahra Assar 2015] [zadeh
87.5%	Support Vector Machine	[Ebenezer O. Olaniyi 2015]
89%	GA and KNN with weights	[Asha G. Karegowda 2014]
85.59%	Feature Selection based Least Square Twin Support Vector Machine	[Divya Tomar 2014]
85.9%	Hybrid Naïve Bayes Classifier and KNN	[Elma Z. Ferdousy 2013]
62.22%	K means	[Shadi I. Abudalfa 2013]
75.15%	GA to reduce feature and determine centers for K-means	[Asha G. Karegowda 2012]
84.44%	KNN	[Muhammad Arif 2012]
86.6667 %	Fuzzy Emphatic Constraints Support Vector Machine	[Mostafa Sabzekar 2010]

2. PROPOSED METHOD

Automated classification for diseases is one of the most important applications that use computers to serve people in health institutions. This study deals with using k-means method in the classification of heart attack and then proposes a method to improve the performance of this method by using the genetic algorithm for reducing properties and delete the insignificant properties**.**

2.1. Classify Database using (K-Means)

Initially it was selected as a method to classify the selected database according to the following steps**:**

Algorithm (k-means) to classify heart attack

Input: global database (Statlog).
Output: accuracy of classification.
Steps:

1. Determine the number of clusters and be 2.
2. Choose two rows of the 270 rows to be the primary centers for the two clusters, and this is done randomly provided that one of the cases is classified (0) while the other is classified as (1).
3. Each case is allocated to the appropriate cluster by calculating the Euclidean distance between the case and the centers.
4. Update the counter responsible for the calculation of the number of cases correctly classified (z) if the k-means status classification is identical to the original category in the database.
5. Update centers by calculating the average values of each cluster.
6. Repeat steps 3-5 if the stop condition is not satisfy, it satisfy when there is no change on the centers and this means that the cases have stabilized in the clusters as a final form.
7. Calculate the final ratio for the classification by the following equation:

$$\text{Rate} = (\text{Number of cases classified correctly} / 270) \times 100\ \% \qquad (1)$$

2.2. Improved Performance of K-Means by the Genetic Algorithm

The classification system depends on properties have a significant impact on the accuracy of system especially some of these properties are not necessary and may cause the system to fall down so it is best to delete them. Because is complex and it is difficult to determine these properties that negatively affect on the performance of the system, this task was assigned to the genetic algorithm.

The genetic algorithm suggests the best properties that k-means can rely on it in the process of classification by using genetic processes to create generations of chromosomes. The proposed properties are derived from the chromosome which is evaluated by running the k-means and calculating the accuracy of the system. After producing several generations the algorithm ends with choosing the chromosome which provides the properties capable of raising the accuracy of the system to the highest possible level. Details of the proposed method are illustrated in the following steps:

Step one: “constructing a genetic foundation”
This phase includes three sub steps:
1) "Specify genetic algorithm coefficients "

The database is stored in an excel file. The file is converted into a two-dimensional matrix containing 270 rows and 14 columns to prevent any errors or changed may be happened on these values and for ease of use. In this step specify some of the parameters that the genetic algorithm are need and as follows:

1. Length of chromosome= number of features in database=13 (where each gene from chromosome is assigned to each property in the database and feature NO. 14 is excluded because it is an ideal output that is used to compare with the system outputs).
2. Number of chromosome in the generation = 50.
3. Number of generations that are created= 60.
4. Probability of crossover = 0.8.
5. Probability of mutation = 0.2.

It is worth mentioning that all the above parameters leave their value to the designer of the algorithm through experiment except chromosome length it is constant because it depends on the number of properties in the database.

2) "Generate primary society"

The initial society is generated randomly according to the parameters specified in the previous step. The output of this step is a generation containing 50 chromosomes. The genes of the chromosome are given binary values (0, 1). If the value of the gene is 0 the feature will be neglected and considered an unnecessary feature to be disposed of. If the value of the gene is (1) this feature is important and is taken into account as one of the features which k-means is based in the classification. For example assume the genetic algorithm generated the next chromosome:

1	0	0	1	1	1	0	0	0	0	0	0	1

This chromosome explains in the following way:

a. Neglected features numbers (not necessary) : 2, 3, 7, 8, 9, 10, 11, 12
b. Important property numbers (proposed) : 1, 4, 5, 6,13

3) "Evaluated chromosome and calculate fitness"

In order to measure the quality of the features proposed by the genetic algorithm the k-means algorithm described in paragraph (3-1) is applied as if the database contained only the features proposed by the genetic algorithm and the other (non-important) features would be disregarded another statement : for each chromosome in the generation a k-means function is called for its evaluation thus the fitness value of the chromosome is the accuracy of the classification calculated by k-means which is illustrated by Equation (1).

Step two: “Great generation through operations genetic”

The genetic algorithm does not stop at the primary generation but continues to generate other generations by simulating the human way of generating backward generations to sustain life. The process of creating a family in human societies begins with the choice of two individuals. This choice is often made randomly and then children are born after marriage. In these children there may be genetic mutations to add diversity in society. This is exactly what the genetic algorithm does during the generation of other generations: selection, crossover, study of the probability of a mutation. The methods used to carry out genetic processes in this research are:

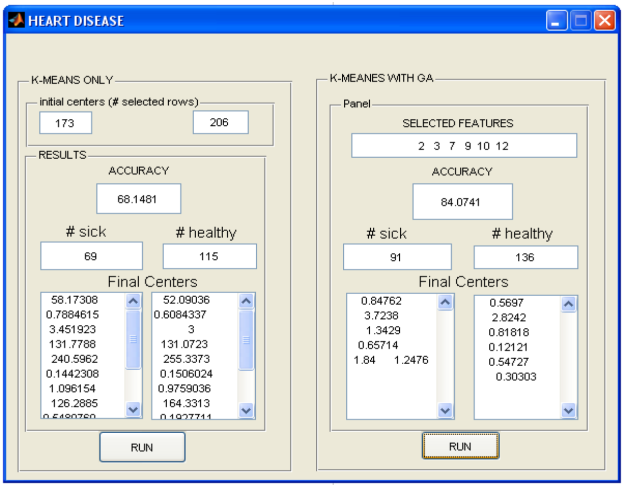
a. Execute selection process by using binary set method.
b. Select uniform mating method to perform crossover.
c. The mutation is implemented in (2m).

As with individual of the primary society the same method is used to evaluate the chromosomes of new generations by calculating the accuracy of the classification by calling the classifier (k-means) as described in step 2. The genetic algorithm continues to generate communities until the number of generations generated reach 60 and the stopping condition adopted in this research.

2.3. Result of Proposed Method and Analysis Performance System

The proposed method was programmed using Matlab version (R2011a). Figure 1 shows the system interface. The interface is designed to compare the performance of k-means alone with the performance of the proposed method to improve classifier k-means when adding the genetic algorithm to select important and useful properties through the following points**:**

a. Display the accuracy of the systems (k-means) and improved method which are calculated by applying Equation (1).
b. Display the number of valid cases (not patient) that were classified by both systems correctly.
c. Calculate the number of infected cases (patient) that both systems can correctly classification.
d. The final values of cluster centers.
e. The system calculates some outputs that are unique for each method such as the primary centers which are the row numbers that are selected to k-means method and also display the important properties discovered by the proposed improved method.


HEART DISEASE

K-MEANS ONLY

initial centers (# selected rows): 173 206

RESULTS

ACCURACY: 68.1481

sick: 69
healthy: 115

Final Centers

```
58.17308     52.09036
0.7884615    0.6084337
3.451923     3
131.7788     131.0723
240.5962     255.3373
0.1442308    0.1506024
1.096154     0.9759036
126.2885     164.3313
```

K-MEANES WITH GA

Panel

SELECTED FEATURES: 2 3 7 9 10 12

ACCURACY: 84.0741

sick: 91
healthy: 136

Final Centers

```
0.84762          0.5697
3.7238           2.8242
1.3429           0.81818
0.65714          0.12121
1.84   1.2476    0.54727
                 0.30303
```

Figure 1. Results of proposed system and comparison with K- means

It is clear that the proposed method which is an improvement for the k-means method using the genetic algorithm gave better results by reducing the insignificant properties in the classification on the contrary the presence of such characteristics reduces the accuracy of the system and thus relied only on the six properties as shown in Figure 1. As a result the system's ability to distinguish healthy cases and cases of this disease are increased which in turn led to an increase in the accuracy of the classification. Table 2 summarizes the results of the system. Figure 2 shows the clear difference between the normal and hybrid methods in terms of accuracy.

Table 2. Result performance of (K-means) and (Genetic K-means)

Properties of method	K-means	GA-(K-means)
Accuracy	68.1481	84.0741
Number of valid cases classified correctly	115	136
Number of infected cases classified correctly	69	91
Number of properties approved in the classification	13	6
Number of selected cases as primary centers	206, 173	204,226 *

* These centers are used only with the six characteristics shown in Figure (1)

Returning to the results of the researchers in the classify of this disease based on computer technology which was explained in Table 1 find that the proposed system obtained good results and acceptable compared to those research as shown in Table 3. Table 4 shows the good performance of the proposed method when compared with research results using the same method.

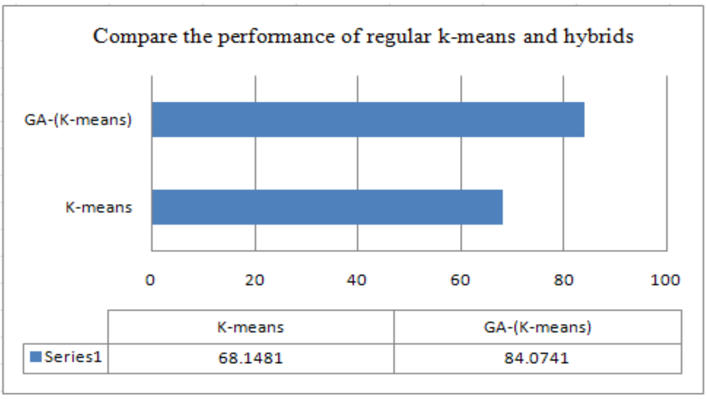


Figure 2. Comparison between normal and hybrid methods in terms of accuracy

Table 3. Compare the proposed system with previous research

Year of Publication	Methods used	Researcher Name	accuracy
2017	Naïve Bayes and Laplace smoothing	Vincy Cherian	86%
2017	Bayes Net	Bharti Dansena	83.5165%
2015	Chaotic Particle Swarm Optimization	Zahra Assar zadeh	75.8889 %
2015	Support Vector Machine	Ebenezer O. Olaniyi	87.5%
2014	GA and KNN with weights	Asha G. Karegowda	89%
2014	Feature Selection based Least Square Twin Support Vector Machine	Divya Tomar	85.59%
2013	Hybrid Naïve Bayes Classifier and KNN	Elma Z. Ferdousy	85.9%
2013	K means	Shadi I. Abudalfa	62.22%
2012	GA to reduce feature and determine centers for K-means	Asha G. Karegowda	75.15%
2012	KNN	Muhammad Arif	84.44%
2010	Fuzzy Emphatic Constraints Support Vector Machine	Mostafa Sabzekar	86.6667 %
	84.0741%	The method suggested in this research	

Table 4. Compare results of system with researches adopted same methods in this research

Year of Publication	Methods used	Researcher Name	accuracy
2013	K means	Shadi I. Abudalfa	62.22%
2012	GA to reduce feature and determine centers for K-means	Asha G. Karegowda	75.15%
2017	GA to reduce features for k-means classifier	Proposed System	84.0741%

It is noted that the results of the program outperform the results of the research referred to in Table 4. It is worth mentioning that the k-means method used in the research surpassed that used by researcher Shadi Abu Delafah in a research published in 2013 where the result of his method is capable of classification by 62% while k-means in this study was able to classify the disease with up to 68% accuracy.

3. CONCLUSIONS

This research discusses the classification of the internationally database known (Statlog) which is related to heart attack using the method (K-means). The accuracy of the classification based on this method was (68%) and then added the genetic algorithm to strengthen the performance of (k-means) by reducing the characteristics adopted during the classification process and found that the genetic algorithm has been instrumental in raising the accuracy of the system where reaching 84% after it was (68%). The results of the application of the system which is designed to classification database cases automatically based on the intelligent recruitment of the computer capabilities without resorting to specialized expertise and comparing the results of this work with the results of the previous works listed in Table 1 the method obtained very good results.

REFERENCES

[1] Vincy Cherian and Bindu M.S "Heart Disease Prediction Using Naïve Bayes Algorithm and Laplace Smoothing Technique", *International Journal of Computer Science Trends and Technology (IJCST)*, vol. 5, no. 2, Mar – Apr 2017.

[2] Bharti Dansena and Amit Kumar Dewangan, "Classification of Heart Disease Using Various Classification Techniques", *International Journal for Research in Applied Science & Engineering Technology (IJRASET)*, vol. 5, no. 5, May 2017.

[3] Zahra Assarzadeh and Ahmad Reza Naghsh-Nilchi,"Chaotic Particle Swarm Optimization with Mutation for Classification", *J Med Signals Sens*, 2015.

[4] Ebenezer O. Olaniyi and Oyebade K. Oyedotun, "Heart Diseases Diagnosis Using Neural Networks Arbitration", *Intelligent Systems and Applications*, 2015.

[5] Asha Gowda Karegowda, "Enhancing Performance of KNN Classifier by Means of Genetic Algorithm and Particle Swarm Optimization", *International Journal of Advance Foundation and Research in Computer (IJAFRC)*, vol. 1, no. 5, 2014.

[6] Divya Tomar and Sonali Agarwal, "Feature Selection based Least Square Twin Support Vector Machine for Diagnosis of Heart Disease", *International Journal of Bio-Science and Bio-Technology*, vol. 6, No. 2, 2014.

[7] Elma Z. Ferdousy, "Combination of Naïve Bayes Classifier and KNearest Neighbor (cNK) in the Classification Based Predictive Models", *Computer and Information Science*, vol. 6, no. 3, 2013.

[8] Shadi I. Abudalfa and Mohammad Mikki, "K-means algorithm with a novel distance measure", *Turkish Journal of Electrical Engineering & Computer Sciences*, 2013.

[9] Asha Gowda Karegowda and et al.,"Genetic Algorithm based Dimensionaliy Reduction for Improving Performance of K-Means Clustering: A Case Study for Categorization of Medical Dataset", *International Journal of Soft Computing*, 2012.

[10] Muhammad Arif and Saleh Basalamah, "Similarity-Dissimilarity Plot For High Dimensional Data Of Different Attribute Types In Biomedical Datasets", International Journal of Innovative Computing, Information and Control, vol. 8, no. 2, 2012.

[11] Mostafa Sabzekar and et al., "Emphatic Constraints Support Vector Machine", *International Journal of Computer and Electrical Engineering*, vol. 2, no. 2, 2010.

BIOGRAPHY OF AUTHOR

Asraa Abdullah Hussein. I got a degree Bachelor of Computer Science from the University of Babylon \ Science Collage for Women \ Department of Computer 2006 High-a good grade, then earned a master's degree from the University of Babylon \ College of Sciences \ Department of Computer Year 2013 High-a good grade in the field of artificial intelligence, and do, Assistant Lecturer at the University of Babylon \ College of Science for women \ Computing Department since 2007 till now.