Improving The Accuracy Of KNN Classifier For Heart Attack Using Genetic Algorithm

Noor Kadhim Ayoob
Noor.kadhum@gmail.com
University of Babylon
Science Collage For Women / Computer Science Department

Abstract

The automatic diagnosis of the diseases using the computer is a fertile field for many researchers who are trying to design systems that help to reduce the mistakes made by inexperienced doctors or because of the influence of the pressures of life .This search deals with the use of (KNN) to diagnose the heart attack and then propose to improve the performance of KNN by using the genetic algorithm to control the basic joints of this method through determining the value of K, database segmentation, in addition to the reducing of the features. The proposed system has succeeded in increasing the accuracy of diagnosis from 75% to 100 %.

Keywords: K nearest, Genetic Algorithm, heart attack, automatic diagnosis.

الخلاصة

التشخيص الأوتوماتيكي للأمراض يشكل حقلا خصبا للعديد من الباحثين الذين يحاولون تصميم نظم تساعد في تقليل الأخطاء التي قد يرتكبها أطباء لا يملكون قدرا كافيا من الخبرة او تحت تأثير ضغوطات الحياة. يتناول هذا البحث استخدام طريقة المجاور الأقرب (KNN) في تشخيص مرض النوبة القلبية و من ثم يقترح تحسين أداء هذه الطريقة عن طريق استخدام الخوارزمية الجينية للتحكم بالمفاصل الأساسية لطريقة المجاور الأقرب وهي تحديد قيمه K وتقسيم قاعدة البيانات إضافة إلى عملية تقليص الخصائص . النظام المقترح نجح في زيادة الدقة إلى 100 % بعد أن كانت 75 % فقط.

الكلمات المفتاحية: المجاور الأقرب، الخوارزمية الجينية، النوبة القلبية، التشخيص الأوتوماتيكي.

1. Introduction

The heart can be affected by different types of dangerous diseases Impact on human lives [1]. According to the World Health Organization (WHO), 12 million deaths occur worldwide every year and heart diseases are the reason [2]. Data mining is applied in medical field to predict diseases [3] such as diseases of the heart, and various types of cancer using medical datasets based on information collected from realistic people.

Computerized diagnostic uses medical databases to classify them in order to build systems that are capable of diagnosing and addressing the emerging situations. There are many techniques that are used in the classification, some of them belong to a category of supervised learning while the unsupervised class includes other techniques. The supervised methods like neural networks or Bayesian classifier use Information extracted from the training data to classify another set of data dedicated to the testing stage. K nearest neighbor (KNN) is another example of supervised methods but it does not extract any information from training set, it just use this set to

make a comparison with a case that is waiting for classification and this is the so-called lazy learning [4].

The number of features in the database (the symptoms of the disease) is also affects the classification process. Depending on the method used, this number can result in time consumption and low accuracy. To get rid of the negative aspects of a the large number of symptoms, a reduction method can be used to choose the symptoms which have a substantial influence in the diagnosis and neglect unimportant symptoms[4], for this purpose genetic algorithm (GA) is the good candidate to do so. This work is dedicated to study how to enhance the diagnosis of heart attack based on a Statlog database.

The paper is organized as follows: section 2 sheds the spot on the most important papers that is work on the same dataset . K-nearest method is presented in section 3. Genetic algorithm principles are reviewed in section 4.Details of the proposed method are described in section 5. Section 6 contains experimental results. Conclusions are presented in section 7.

2. Related work

As stated in section 1, the Statlog dataset used in researches that are designed to diagnose heart attack, Statlog can be obtained from the UCI Repository. The total number of instances in the database is (270) with (13) symptoms, the feature number 14 represents the final diagnosis .

The features include : age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved, exercise induced angina, oldpeak, the slope of the peak exercise st segment, number of major vessels and Thal [5] [6]. A number of papers that try to diagnose this disease using Statlog dataset is summarized in Table 1. The researches are arranged by year of publication.

Table 1: The summary earlier work

Researcher name	year	method	Accuracy
Zahra Assarzadeh [7]	2015	Chaotic Particle Swarm Optimization	75.8889 %
Ebenezer O. Olaniyi[6]	2015	Support Vector Machine	87.5%
Asha G. Karegowda [4]	2014	GA and KNN with weights	89%
Divya Tomar[8]	2014	Feature Selection based Least Square Twin Support Vector Machine	85.59%
Elma Z. Ferdousy[5]	2013	Hybrid Naïve Bayes Classifier and KNN	85.9%
Shadi I. Abudalfa[9]	2013	K means	62.22%
Asha G. Karegowda [10]	2012	GA to reduce feature and determine centers for Kmeans	75.15%
Muhammad Arif [11]	2012	KNN	84.44%
Mostafa Sabzekar[12]	2010	Fuzzy Emphatic Constraints Support Vector Machine	86.6667 %
Essam Al-Daoud [13]	2009	Three types of neural net works (LVQ, LM, GRNN)	88.8%

It is clear that over the years, the research used different methods to diagnose a heart attack, but the performance of most of these methods is located within the range of the 80's %, if not less. Such a dataset constitutes a challenge for any researcher interested in these fields and this is the reason for choosing the Statlog database in this study.

3. K-nearest technique

The K- nearest neighbor classifier is a supervised classifier that has a good performance [14].As any supervised learning algorithm, KNN algorithm passes through two phases : training phase and test phase [15].The steps of KNN is explained by the following algorithm:

Algorithm1: KNN Classifier

```
For i=1 to No. of cases in test set
 Begin
   -Assign the current test case to the vector Y.
  - For j= 1 to No. of cases in train set
    Begin
        - Assign the current train case to the vector X.
        - Compute the distance between X and Y.
        - Store the result in a data structure.
    End
  -Select First K samples with minimum distance to test case.
  -Count the frequency of each class in the selected K samples.
  -Choose the class with the larger frequency to be the class of the test case.
 End
```

Several methods can be used to measure the distance, the most common one is Euclidean distance.

4. Genetic algorithm

John Holland founded the rules of genetic algorithm[16] in 1975. A genetic algorithm is one of the most popular optimization tools in the evolutionary algorithms family. Genetic algorithm relies on generating a set of the nominated solutions - chromosomes - and then pick the best solution.

The method starts with randomly generating chromosomes [17] called population, and then iterative steps are implementing to create a new set of solutions. The iterative steps are mainly: selecting two chromosomes to be parents, mating them to have two offspring, and finally performing a mutation if the probability allow [18].For each step there are number of methods to implement. For example, the selection of binary tournament is a powerful method of selecting two individuals at random, in preparation for mating stage. In addition to this method it is possible to use the selection of the roulette wheel or rank etc.

After the selection steps, the genetic algorithm moves toward another phase, a phase of mating , to generate two offspring by mixing the information of parents in some way, This can be carried out using traditional methods such as 1X, 2X or UX

[19], but the nature of some problems may impose certain requirements and therefore need specific types of mating like PMX,OX,CX, etc.

The mutation process is applied on the individuals produces by the previous step under low probability to maintain the stability of population. Replacing the values of two genes is a well-known mutation method. If the chromosomes are encoded as binary values, the complement approach can be used. Complements means that the value of selected gen is flipped form 0 to 1, or become 1 if it was 0 [20]. One way to rise performance of the genetic algorithm up is to preserve the n best individual in the old population and uses them as substituent for the n worst individuals in the new population. This cycle is continued till the stop criterion is met by reaching a specific number of generations, a goal is satisfied, or no change in the performance is observed.

5. Using genetic algorithm as KNN enhancer

This section is dedicated to clarify the parameters of GA used in this work but first, performance when using classic KNN classifier with the Statlog database is presented. The results of classic KNN classifier are summarized in table 2. Obviously, even when 97% from data is devoted for training, the results are not satisfied.

Table 2: classic KNN results

K	50/50	60/40	70/30	80/20	90/10	97/3
1	55.555	55.555	57.790	51.851	66.666	75
3	63.703	57.407	62.963	62.963	59.259	75
5	66.666	67.592	67.901	70.370	66.666	75
7	68.888	69.444	69.135	68.518	59.259	62.5
9	65.925	64.814	69.135	64.814	62.963	62.5
11	67.407	66.666	69.135	64.814	55.555	75
13	68.888	71.296	69.135	62.963	55.555	62.5
Best	**68.888**	**71.296**	**69.135**	**70.370**	**66.666**	**75**

The classic KNN classifier is based on measuring the similarity among database instances. Since this method does not give the desired accuracy, it means that the data is heterogeneous because there are a few similarities among the cases. At this point the idea of looking for similarity among a portion of the properties and not all seems to be a promising idea and this task is assigned to GA.

The role of the genetic algorithm in this search is to find a partial similarity among the cases of database by selecting the most important properties among the thirteen features. It is clear that the accuracy significantly affected by the value of K and the size of training and testing sets, so there are additional tasks carried out by the GA: the proposal of an appropriate value for K and the data segmentation into train set and test set. The parameters of genetic algorithm that are used in this search are:-

1- Chromosome structure:

In order to make the idea clear to the mind, the chromosome is divided into three parts:

(a) The first gene: responsible for determining the value of K which is an odd number between 1 and 13.
(b) The second gene: specifies how to divide the database and the value of this gene is also integer within the range of 50 to 100. For example, if the value of this gene (70) , 70% of the database is allocated for training set and the remaining (30 %) is devoted to test set.
(c) Genes from 3 to 15 are binary values devoted for the features so that there is a gene for each feature. These genes define which of the 13 features involved in the partial similarity measurement. The gene of the value 1 is selected to take part in the similarity measuring while the gene of 0 value is neglected. Based on the foregoing, the length of chromosome is 15 (1 for K plus 1 for partitioning data plus 13 for features). For example:

3	87	0	1	1	0	0	0	0	0	1	1	0	1	1

The system interprets the previous chromosome as follows:
(a) K=3(depends on the closest three cases to decide if the patient is infected or not.)
(b) The size of training set = (87/ 100) × 270 = 234.9 = 235
(c) The size of test set= 270-208= 35
(d) The features that are selected: 2, 3 , 9 ,10 ,12 and 13. i.e. the similarity among the instances of database is measured depending on these features only while features (1, 4 ,5, 6, 7, 8 and 11) are discarded. This chromosome is able to classify the data by 94.2857%.

2- Fitness function:

Each solution (chromosome) detected by GA is evaluated by KNN classifier to find the quality of the solution which is simply obtained by calculating the percentage of cases that are classified correctly.

3- Other parameters:

As seen in the section 4, the genetic algorithm passes through a number of steps and for each step there are some methods for implementing. Table 3 summarizes this information:

Table 3: important genetic parameters

Process	Method
Mating type	UX
Mutation type	complement (Applied on genes 3 to 15 only)
Selection method	Binary tournament
population size	50
Stopping criterion	stopped when the No. of generations is 25

The genetic algorithm uses the previous parameters to find the subset of significant feature for KNN and this process go through the following steps:

Algorithm2: GA-KNN Classifier

1. Initialize the population randomly .
2. For each chromosome in the initial population, apply K–Nearest Neighbor exploiting information extracted from the chromosome and use significant features to find the classification accuracy as fitness of the chromosome.
3. Repeat the steps a-d until terminating condition (maximum number of generations) is reached.
   a. Select two chromosomes using binary tournament.
   b. Apply uniform crossover (UX) operation on the selected individuals.
   c. Apply complement mutation by selecting any bit randomly and flip its value.
   d. Calculate the fitness of new individuals in the same manner mentioned in 2
4. Select the chromosome resulting in highest classification accuracy of KNN classifier to be the solution presented by the system.

The KNN classifier in this algorithm is the same as described in the algorithm 1 except that when it compares the cases of training and testing, it does not take all the thirteen features, instead, it compares the features proposed by the chromosome those with the value 1, for example, Suppose that genetic algorithm explores the following chromosome and get ready to evaluate its quality:

7	69	0	0	0	0	0	1	0	0	1	0	0	0	0

After interpreting this chromosome, the system concludes that:

-K=7

-The size of training set = (69/ 100) × 270 = 186 instances for training.

-The size of test set= 270 – 186 = 84 instances for test.

-The selected features are : 6 and 9.

Let's see the difference between the work of classic KNN classifier and KNN supported by GA technique when measuring the similarity between two cases such:

Training instance (X)

70	1	4	130	322	0	2	109	0	2.4	2	3	3

Test instance(Y)

67	1	4	160	286	0	2	108	1	1.5	2	3	3

A)Classic KNN : Take into account all the thirteen symptoms to calculate the distance between the two cases X andY.Using Euclidean distance, the distance (d) is calculated as follows:

$$d(X,Y) = \text{sqrt}\,(\text{sum}[(X-Y)^2]) = 46.9873.$$

B)GA-KNN

A quick look to the data of these cases gives a feeling that the two are somewhat similar but the distance scale shows the opposite in the classic KNN. In the proposed system, the distance between the two cases is calculated depending on the selected features that are founded by genetic algorithm (feature no. 6 and feature no. 9 only):

- The value of the feature no.(6) is (0) in both cases.
- The value of the feature no.(9) is (0) in training instance and (1) in the test instance.
- $d(X,Y)= \text{sqrt} ((0-0)^2 + (1-0)^2) = 1$ .This distance pointing to the fact that the similarity between the two cases is high because it is close to zero.

6. Results

This section describes the results obtained by the proposed system. To make the comparison between classic KNN and GA-KNN honest and to be able to see the impact of genetic algorithm, the same segmentation of the dataset is used to get the outputs which are recorded in table 4.

The results presenting in the table shows that the minimum accuracy is (85.9 %) with three important features and the best performance is (100%) when semi-full training(97% of data as training set) is used, while the performance classic KNN is 75% using the same segmentation as seen in table 2. This draws attention to the fact that the credit for the accuracy achieved by the new system is not due to the use of large size of the training set, the reason for perfect performance is the coalition of the three factors which their values determined by GA. A quick sight on the results of the two classifiers is summarized in table 5.The same accuracy can be obtained using the full training with k=1 and features no. 1, 2 ,4, 10, and 13 as shown in table 6.

Table 4: The primary results of GA-KNN classifier

K	50/50		60/40		70/30		80/20		90/10		97/3	
	Result	No. of Feature	result	No. of feature	result	No. of feature	result	No. of feature	result	No. of feature	result	No. of feature
1	82.22	6	84.26	6	86.42	5	87.04	6	92.59	5	87.5	6
3	85.19	6	86.11	6	90.12	5	87.04	4	96.30	5	100	4
5	85.93	6	87.96	6	86.42	6	87.04	3	92.59	5	100	4
7	85.93	4	84.26	6	86.42	6	88.89	5	88.89	7	100	4
9	85.93	3	85.19	6	88.89	6	88.89	4	85.19	5	100	4
11	84.44	4	85.19	4	86.42	5	88.89	4	85.19	4	100	7
13	85.93	3	85.19	4	87.65	3	87.04	4	88.89	7	100	5
Best	**85.93**		**87.96**		**90.12**		**88.89**		**96. 30**		**100**	

The good news is that the system is not only able to diagnose the disease with accuracy of 100%, but also it can do so with very few properties where only two features out of 13 are needed in order to obtain this rate as it is shown in table 6.

Table 5: A comparison between the performance f KNN with and without the using of GA

Data / System	**50/50**	**60/40**	**70/30**	**80/20**	**90/10**	**97/3**
Classic KNN	68.89	71.30	69.14	70.37	66.67	75
GA-KNN	85.93	87.96	90.12	88.89	96.30	100

Table 6: Top five results

K	Size of train set	# features	Selected features	Classification rate
5	98%	2	[2, 7]	100%
1	100%	5	[1, 2, 4, 10, 13]	100%
3	97%	4	[5, 10, 11, 12]	100%
3	87%	5	[6, 7, 10, 12, 13]	97.14%
3	88%	5	[6, 7, 10, 12, 13]	96.88%

7. Conclusions and future works

Experimental results demonstrate the efficiency of genetic algorithm in treating poor performance of KNN and its ability to strengthen the performance significantly through the negligence of the properties that are not important in the database and control the segmentation process in addition to finding the value. Tests are showed that the system has the ability to diagnose the disease and classify the database with accuracy of 100% depending on the two properties and only five neighborhoods. In the future, the following options can be applied:

1. Using a local data taken from Iraqi environment.
2. Collecting data for other heart disease and modify the system to be able to diagnose several types of heart disease.
3. Studying the effectiveness of the system using databases that belong to other diseases.

References

[1] Salha M. Alzahani and Etal. " An Overview of Data Mining Techniques Applied for Heart Disease Diagnosis and Prediction ", Lecture Notes on Information Theory, Vol. 2, No. 4, 2014.

[2] Suganya and Tamije Selvy," A Proficient Heart Disease Prediction Method Using Fuzzy-Cart Algorithm", International Journal of Scientific Engineering and Applied Science, Vol. 2, No.1, 2016. [3] Hardik Maniya and Etal., "Comparative study of Naïve Bayes Classifier and KNN for Tuberculosis ", International Conference on Web Services Computing (ICWSC), 2011.

[4] 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.

[5] Elma Z. Ferdousy, "Combination of Naïve Bayes Classifier and K-Nearest Neighbor (cNK) in the Classification Based Predictive Models ", Computer and Information Science; Vol. 6, No. 3; 2013.

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

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

[8] 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.

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

[10] Asha Gowda Karegowda and Etal.,"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.

[11] 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.

[12] Mostafa Sabzekar and Etal., "Emphatic Constraints Support Vector Machine", International Journal of Computer and Electrical Engineering, Vol. 2, No. 2, 2010.

[13] Essam Al-Daoud," Acomparission Between three Neural Network models for Classification Prpblems" Journal of Artificial Intelligence,Vol.2,No.2, 2009.

[14] Hussein A. Lafta and Esraa A. Hussein, " Design a Classification System for Brain Magnetic Resonance Image ", Journal of Babylon University/Pure and Applied Sciences, Vol.21, No.8, 2013.

[15] Abdulrazzaq and ShahrulA. Noah, "Improving the annotation Accuracy of Medical Images in Image CLEFmed 2005 Using K-Nearest Neighbor (k NN) Classifier" ,Asian Journal of Applied Science ,Vol.8,No.1, pp.16-26, 2015.

[16] Naveen Singh and Etal," computational intelligence in circuit synthesis through evolutionary algorithms and particle swarm optimization", International Journal of Advances in Engineering & Technology, Vol. 1,No. 2,pp.198-205, 2011.

[17] Dilbag Singh and Etal. "To Design a Genetic Algorithm for Cryptography to Enhance the Security ", International Journal of Innovations in Engineering & Technology, Vol. 2, No. 2, 2013.

[18] Girdhar Gopal and Etal," Enhanced Order Crossover for Permutation Problems" International Journal of Innovative Research in Science, Engineering and Technology, Vol. 4, No. 2, 2015.

[19] Jorge Magalhães-Mendes," A Comparative Study of Crossover Operators for Genetic Algorithms to Solve the Job Shop Scheduling Problem ", Wseas Transactions on Computers, Vol. 12, No.4, 2013.

[20] Pankaj Mehta and Etal. " Genetic algorithm & Operators ", International Journal Of Engineering Sciences & Research Technology, Vol.4,No.2, 2015.