International Journal of Bio-Science and Bio-Technology
Vol.8, No.2 (2016), pp.363-372
http://dx.doi.org/10.14257/ijbsbt.2016.8.2.34

A Computational Intelligence Method for Effective Diagnosis of Heart Disease using Genetic Algorithm

P. Siva Kumar[1], D. Anand[1], V. Uday Kumar[1], Debnath Bhattacharyya[2] and Tai-hoon Kim[3*]

[1]Department of Computer Science and Engineering, K L University, Andhra Pradesh
[2]Department of Computer Science and Engineering, College of Engineering, Bharati Vidyapeeth University, Pune-411043, India
[3]Department of Convergence Security, Sungshin Women's University, 249-1, Dongseon-dong 3-ga, Seoul, 136-742, Korea
[1]siva.rise@gmail.com, [1]ananddama92@gmail.com, [1]uday009u@gmail.com
[2]debnathb@gmail.com, [3]taihoon@daum.net

Abstract

In recent years improvement of new and effective medical domain applications has vital role in research. Computational Intelligence Systems (CIS) has profound influence in the enlargement of these effective medical field applications and tools. One of the prevalent diseases that world facing is heart disease. The Computational Intelligence Systems uses input clinical data from different knowledge resources throughout the world and applies this data on different computational intelligence tools that uses sophisticated algorithms. The sophisticated algorithms plays prominent role in the construction of medical clinical analysis tools. These tools may be used as an extra aid for the clinical diagnosis of the diseases for the doctors and clinicians. In this paper a novel method for the diagnosis of heart disease has been proposed using Genetic Algorithms. In this approach an effective association rules are inferred using Genetic Algorithm approach which uses tournament selection, crossover, mutation and new proposed fitness function. The Cleaveland data set is used for the experimentation. This data set is collected from the UCI machine learning repository experimental results are prominent when compared with some of the supervised learning techniques.

Keywords*: Computational Intelligence Systems, Heart Disease, supervised learning techniques, Genetic Algorithms*

1. Introduction

The extraction of useful data and mapping of hidden patterns and relationships from the huge databases, we need to merge different technologies. One such is merging the data mining with the statistical analysis, machine learning and database technology [1]. This technology is used in many areas including the medical services. Data mining techniques can be used effectively in surgical procedures, medical tests, medication, and the discovery of relationships among clinical and diagnosis data along with predicting the diseases [2]. The knowledge, rich data that is available in the database is not the one the clinical decisions are made, these are made by the doctors observations and experience [3]. The problem in the decisions is that the doctor's expertise is not even in every sub-specialty and is in several places as a scarce resource [3].

* Corresponding Author

ISSN: 2233-7849 IJBSBT

The technology is more useful when we come across the patients suffer with more than one type of disease of same category. In such cases the information obtained from the patient may be interrelated with the signs and symptoms in the medical diagnosis where the physicians may not be able to diagnose accurately [4]. So the quality services with an affordable cost can be given to the patients with the help of appropriate computer-based information and/or decision support systems [5]. With the help of the clinical decision support systems (CDS) the individuals with the appropriate knowledge and persons specific information can intelligently filter the data and can be presented at appropriate times. A few health care institutions using the CDS effectively in obtaining the information needed for the users [6]. There is a gradual incline in the clinical databases [7] where storage of patients information is done and usage of data mining on this data to obtain the useful information which can lead to a successful decision. Clinical decision process is not an easy task it must be accurate and sometimes quick, but uncertainty occurs in every stage. The problem of uncertainty is due to different reasons which includes patients not describing accurately what has happened to them or how they suffer, doctors and nurses cannot explain exactly what they detect, some degrees of error in the laboratory reports, inability of the medical researchers not precisely characterizing how diseases modify the normal functioning of the body and no one can precisely determine ones prognosis [8], [9]. There was a lot of improvement has been achieved in both theoretical and practical areas since the idea of the computer-based clinical decisions has been started, but still a number of obstacles exist in the implementation of the technology.

In this proposed work Genetic Algorithm technique is used to construct a computational intelligence technique for the diagnosis of heart disease. The preprocessing of the data is done on the heart disease dataset by removing the missing values and the noise information.

2. Heart Disease

In the present scenarios, there is drastic change can be observed in the health care of the world's population where the major deaths are being occurred from the non-communicable diseases (NCDs), such as diabetes, cancer, depression and heart disease, replacing infectious diseases and malnutrition as the leading causes of disability and premature death. Interestingly it is not the developed countries that are affected by the cardiovascular disease but we can find a large number of low income countries including both men and women are equally affected [14, 15].The developing countries are showing an increased rate of CVD which is almost double compared to the developed countries [17]. All over the world cardiovascular diseases account for high morbidity and mortality.

A decline is seen in some countries due to the health policies of those governments. The term "cardiovascular disease" is a category of heart disease comprises of a variety of conditions that upset the heart and the blood vessels and the way in which blood is pumped and circulated in the body. Cardiovascular diseases are contributing towards an ever-increasing proportion of the non-communicable diseases in the developing countries [10-13]. In the countries like United States the American Heart Association in "2015 Heart Disease and Stroke Statistics Update" claimed that Cardiovascular disease claiming the number of deaths 17.3 million per year and expected that it will grow to 23.6 million by 2030. In 2011 there were 787,000 deaths due to heart disease, stroke and CVD in U.S. that is about one in every three deaths [16].

In India 52% of deaths are from CVDs occurring before 70 years compared to 23% in developed countries showing that younger generations are more effected [18].even in rural areas it was found that 32% of deaths are due to CVD where a survey was conducted in 45 villages proving that the affect has reached even to the rural India [19]. Coronary heart disease (CHD) is caused by the decreased blood and oxygen supply to the heart due to the narrowing of the coronary arteries. It also includes the Heart attack that is

myocardial infarctions in scientific terms [20].One should note that several types of cardiovascular disease such as high blood pressure, coronary artery disease, vascular heart disease, stroke, or rheumatic fever/rheumatic heart disease exist.

3. Data Set Description

The data set for this experiment was obtained from the data mining repository of the University of California, Irvine (UCI) [21]. The heart disease data set section contains four databases Cleveland, Hungary, Switzerland and the VA Long Beach. Most of the researchers use the Cleveland data set; hence we also opted for it. Cleveland data set is taken which contains 303 instances with 76 attributes and the data was collected by the Robert Detrano. Previous researches used 14 attributes out of available 76 attributes from the pre-possessed Cleveland data set. The attributes that are considered are Age, sex, chest pain type, resting blood pressure, serum cholesterol in mg/dl, fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved, exercise induced angina, ST depression, slope of the peak exercise ST segment, number of major vessels, thal and diagnosis of heart disease are presented.

4. Previous Related Work

John Holland and his group first introduced the Genetic Algorithm (GA) in early 60s [22]. It was extensively adopted to solve many optimization problems in engineering and science. GA was proved efficient search in complex space. The GA is more useful for large datasets. A hybrid classification algorithm was proposed by D. Kelly, Jr and Lawrence Davis using k-nearest neighbor and GA [23]. De Falco *et. al.,* used evolutionary system for inferring explicit classification rules from breast cancer data set [24]. Korkut Koray Gundogan *et. al.*, used GA technique with non-random initial population and uniform operator method to infer classification rules [25]. Dehuri and Mall proposed a multi-objective genetic algorithm for inferring highly predictive and comprehensible classification rules on massive datasets.

5. Proposed Methodology for Heart Disease Classification

The initial Genetic Algorithm proposed by J H Holland was based on “Survival of the fittest”. In this paper the heart disease data set is classified using the GA classifier. It uses association rule mining concept to infer best and effective rules from the given input data set. The initial rules which are known as chromosome is represented using the combination of attributes and operators. The individual rules consist of two parts that is ‘if part’ and ‘then part’.

5.1. Chromosome Representation

The chromosomes are build using if → then rules. A general chromosome is represented as below.

$$\mathbf{A_i\ OP_i\ V_{ij}}$$

Where, A_i represents i-th predictor attribute,

OP_i represents set of operators { =, <, >, <=, >= } used to construct the if → then rules and V_{ij} represents j-th value of i-th attribute.

$\mathbf{A_i\ OP_i\ V_{ij}}$	…….	$\mathbf{A_k\ OP_k\ V_{kj}}$

Figure 1. Individual Genome Representation

Attribute-1	Attribute-2	Attribute-3	Class Label

Figure 2. Rule with Attributes and Class Label

=	>=	<=

Figure 3. Operator Combination for the above Attributes

The absence of attribute is represented with symbol "#" in the operator combination. The general rules are as follows

P → Q

P ^ Q value represents the number of samples in the data set that are fulfilling the antecedent and consequent parts of the given rule.

P represents the number of samples in the data set that are satisfying only the antecedent part in the rule.

Q represents the number of samples that are satisfying the only the consequent part in the rule.

L is used to specify the possible maximum number of attributes that are participating in the rule.

B represents number of attributes that are participating in the current rule.

The general performance measures True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) are used to construct the fitness function [26].

Sensitivity: The ability of the method to identify the occurrence of target class accurately.

$$\text{Sensitivity} = TP/[TP+FN]$$

Specificity: The ability of the method to separate the target class accurately.

$$\text{Specificity} = TN/[TN+FP]$$

Comprehensibility: It is measured using the number of attributes on the left hand side of the rule [V6].

$$\text{Comprehensibility} = [L-B]/[L-1]$$

$$\text{Fitness Function} = \text{Sensitivity} * \text{Specificity} + (0.2) * \text{Comprehensibility} \quad [27].$$

The flow chart for the proposed GA is presented in Figure 5. It contains those steps as follows:

Step 1: Generation of Initial population

Step 2: Calculate the fitness values of each individual

Step 3: Select individuals to generate next population

Step 4: Apply cross over operation to generate next population

Step 5: Apply mutation operator to generate better siblings

Step 6: Continue above five steps until it met stopping condition.

Step 7: Sort the rules based on their fitness values.

Step 8: Measure the classifier performance using test set on inferred rules.

In this paper tournament selection method is used for selecting the offspring individuals. This predicts the convergence rate of the gap between individuals.

5.2. Tournament Selection Steps

Step 1. Choose any two rules randomly from first half (pool 1) and second half (pool 2) of the total population where one is taken from pool-1 and other is taken from pool-2.
Step 2. Choose the parents with maximum fitness value from pool 1 and pool 2 to create pool 3.
Step 3. Apply GA operators on pool 3 rules.

5.3. Crossover Operation

In crossover operation it takes two individual rules as input and selects a random point in the rules and interactions sub expression at the rear point. Types of crossover operators are used depending up on the selection of random point on the rule that is single crossover, double crossover and multi crossover; it is shown in the Figure 4.

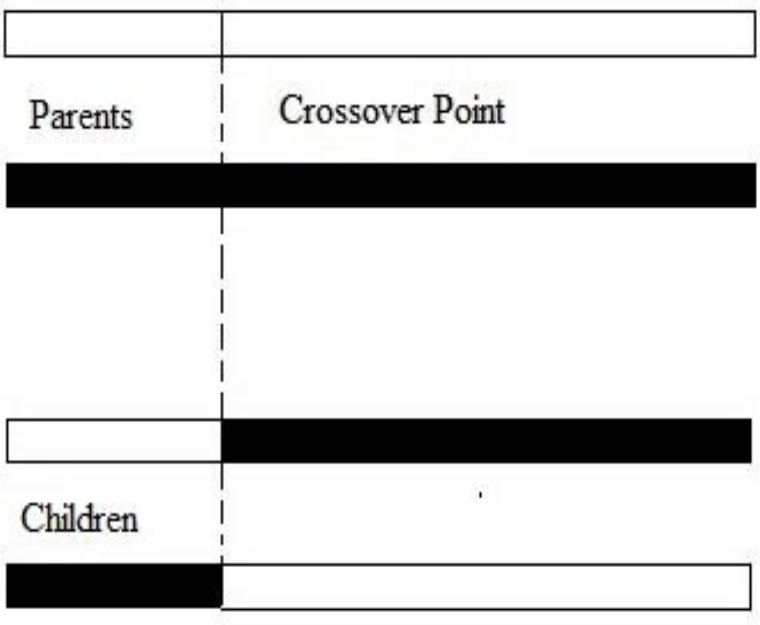

Figure 4. Single Point Crossover

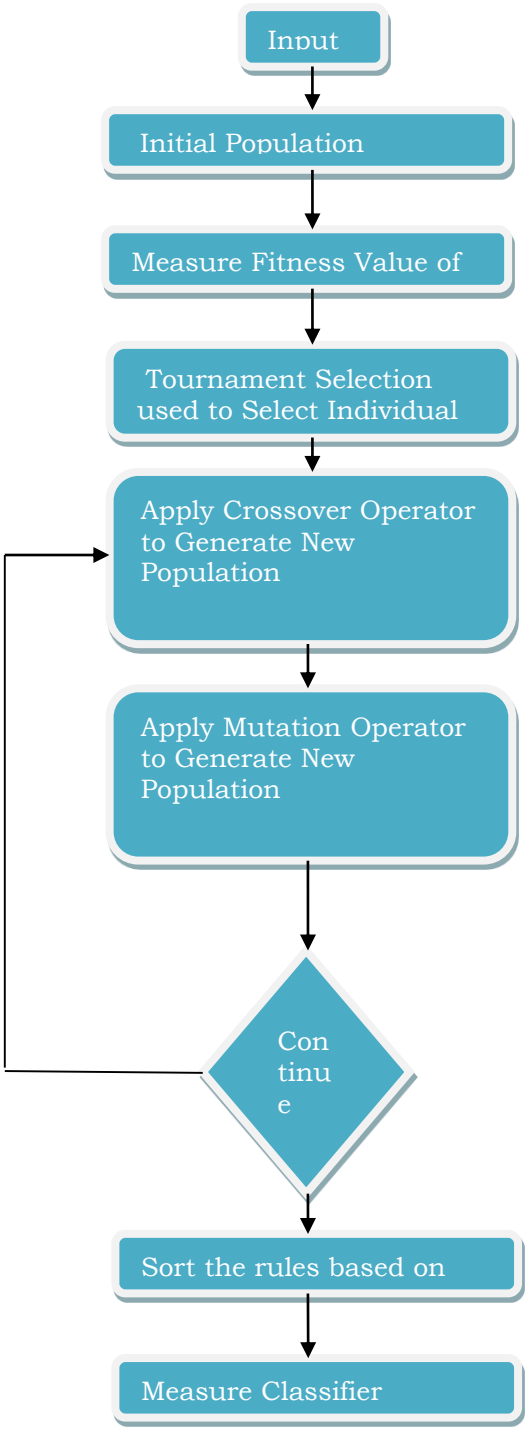

Figure 5. Genetic Algorithm Flow Chart

5.4. Mutation Operator

The genetic diversity from one offspring to another is continued using one of the genetic algorithm operator known as mutation operator. Mutation changes one or more gene values in the chromosome from its initial state. This mutation operator may produce better offspring by making changes in the values of the parent chromosome. In this method the attribute values in the rule are mutated and applied to produce new off springs with better fitness values.

5.5. Stopping Criteria

If all the instances in the data set belongs to the same class label
If data set have all related instances in the attribute list
If the data set is empty

5.6. Performance Metrics

Accuracy, Sensitivity and Specificity are the most widely used performance metrics in computation intelligence systems [28]. k-fold cross validation is used to improve the measures of performance. This work has been carried out using the 3 fold cross validation approach. The performance metrics are presented in table 1.

Table 1. Performance Measures.

Measure	Formula
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$
Sensitivity	$\frac{TP}{TP+FN}$
Specificity	$\frac{TN}{TN+FP}$

6. Results and Conclusion

In this work the authors described GA technique to construct computational intelligence methods for the diagnosis of heart disease. The performance of the model is validated using 3-fold cross validation approach. The input heart disease data set is taken from UCIML repository. The average accuracy, sensitivity and specificity obtained each fold is presented in the tables 2,3 and 4. The average rules that are inferred to classify the test data instances in each fold are around 8. The better and consistent results can be obtained using hybrid models that are by combining GA with other well suited techniques. This approach can be used for other disease datasets.

Table 2. Presentation of Accuracy, Sensitivity and Specificity Using Fold-1

Initial Rules	Fold-1		
	Accuracy	Sensitivity	Specificity
25	83.6	68.1	79.8
50	87.7	69.4	78.1
75	90.6	77.3	80.9

Table 3. Presentation of Accuracy, Sensitivity and Specificity Using Fold-2

Initial Rules	Fold-2		
	Accuracy	Sensitivity	Specificity
25	79.1	65.4	78.3
50	85.6	70.1	78.2
75	89.7	75.2	80.1

Table 4. Presentation of Accuracy, Sensitivity and Specificity Using Fold-3

Initial Rules	Fold-3		
	Accuracy	Sensitivity	Specificity
25	82.8	69.2	76.8
50	87.2	70.4	80.1
75	91.1	79.0	81.1

Table 5. Average Accuracy with Different Initial Rules

Initial Rules	Average Accuracy (%)
25	81.83
50	86.83
75	90.46

The performance of the proposed classifier has been tested by comparing with state-of-art techniques which is presented in the table 6. When compared with the other models the proposed method gives better accuracy over the other methods on Cleveland data set.

Table 6. Average Accuracy Obtained with Different Classifiers Using Cleveland Data Set

Method	Average Accuracy (%)	Authors
28-NN, stand Euclidean	85.1	WD/KG
LDA	84.5	Ster & Dobnikar
Fisher discriminant analysis	84.2	Ster & Dobnikar
16-NN	84	NCU
25-NN	83.6	NCU
FSM	82.4	R.Adamczak
Naïve Bayes	82.5-83.4	Rafal, Ster, Dobnikar
C4.5	77.8	Bennet & Blue
Our method	**90.46**	--

References

[1] T. Bhavani, "A primer for understanding and applying data mining", IT Professional, vol. 2, no. 1 **(2000)**, pp. 28-31.
[2] T. Tzung-I, G. Zheng, Y. Huang, G. Shu and P. Wang., "A comparative study of medical data classification methods based on decision tree and system reconstruction analysis", Industrial Engineering and Management Systems, vol. 4, no. 1, **(2005)**, pp. 102-108.
[3] P. Latha and R. Subramanian, "Intelligent heart disease prediction system using CANFIS and genetic algorithm", International Journal of Biological, Biomedical and Medical Sciences, vol. 3, no. 3, **(2008)**.
[4] D. Shanthi, G. Sahoo, and N. Saravanan., "Input feature selection using hybrid neuro-genetic approach in the diagnosis of stroke disease", IJCSNS, vol. 8, no. 12, **(2008)**, pp. 99-107.
[5] P. Sellappan and R. Awang, "Intelligent heart disease prediction system using data mining techniques", In Computer Systems and Applications, 2008. AICCSA 2008. IEEE/ACS International Conference on, pp. 108-115. IEEE, **(2008)**.
[6] M. George, K. James, D. Bader, J. Frantsve-Hawley and K. Aravamudhan. "Clinical decision support chair side tools for evidence-based dental practice", Journal of Evidence Based Dental Practice, vol. 8, no. 3, **(2008)**, pp. 119-132.
[7] A. Gupta, N. Kumar and V. Bhatnagar, "Analysis of Medical Data using Data Mining and Formal Concept Analysis", Proceedings of World Academy Of Science, Engineering Andechnology, vol. 6, no. 3, **(2005)** June, pp. 119-132.

[8] S Peter., "Uncertainty and decisions in medical informatics", Methods of Information in Medicine-Methodik der Information in der Medizin, vol. 34, no. 1 **(1995)**, pp. 111-121.
[9] K Guilan, D.-L. Xu, and J.-B. Yang, "Clinical decision support systems: a review on knowledge representation and inference under uncertainties", International Journal of Computational Intelligence Systems, vol. 1, no. 2, **(2008)**, pp. 159-167.
[10] U. Belgin, J. Alison Critchley and S. Capewell, "Explaining the decline in coronary heart disease mortality in England and Wales between 1981 and 2000", Circulation 109, no. 9, **(2004)**, pp. 1101-1107.
[11] World Health Organization, "Non communicable diseases in South-East Asia region. A profile, New Delhi: profile (2002)", **(2013)**.
[12] R K. Srinath, BShah, C Varghese, and A Ramadoss., "Responding to the threat of chronic diseases in India", The Lancet , Vol. 366, No. 9498 **(2005)**, pp. 1744-1749.
[13] K, Kari, *et al*., "Estimation of contribution of changes in classic risk factors to trends in coronary-event rates across the WHO MONICA Project populations", The Lancet 355.9205 **(2000)**, pp. 675-687.
[14] World Health Organization. The world health report 2002: reducing risks, promoting healthy life. World Health Organization, (**2002**).
[15] E, Majid, Alan D. Lopez, A Rodgers, St Vander Hoorn, and C JL Murray., "Selected major risk factors and global and regional burden of disease", The Lancet 360, No. 9343 **(2002)**, pp. 1347-1360.
[16] http://circ.ahajournals.org
[17] G Thomas A., "Cardiovascular disease in the developing world and its cost-effective management", Circulation 112, No. 23 **(2005)**, pp. 3547-3553.
[18] R, K. Srinath, and S Yusuf., "Emerging epidemic of cardiovascular disease in developing countries", Circulation 97, No. 6 **(1998)**, pp. 596-601.
[19] J Rohina, *et al*., "Chronic diseases now a leading cause of death in rural India—mortality data from the Andhra Pradesh Rural Health Initiative", International Journal of Epidemiology, Vol. 35, No.6 **(2006)**, pp. 1522-1529.
[20] P Shantakumar B., and Y. S. Kumaraswamy., "Intelligent and effective heart attack prediction system using data mining and artificial neural network", European Journal of Scientific Research, Vol. 31, No. 4 **(2009)**, pp. 642-656.
[21] C. L. Blake, and C J. Merz., "UCI Repository of machine learning databases [http://www. ics. uci. edu/~ mlearn/MLRepository. html]. Irvine, CA: University of California", Department of Information and computer science, Vol. 55 **(1998)**.
[22] J H Holland,., "Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence", U Michigan Press, **1975**.
[23] K Jr, James D., and L Davis., "A Hybrid Genetic Algorithm for Classification", IJCAI, vol. 91, pp. 645-650. **1991**.
[24] I De Falco, , A Della Cioppa, A Iazzetta, and E Tarantino, "An evolutionary approach for automatically extracting intelligible classification rules", Knowledge and Information Systems, vol. 7, no. 2, **(2005)**, pp. 179-201.
[25] G, Korkut Koray, B. Alatas, and A Karci., "Mining classification rules by using genetic algorithms with non-random initial population and uniform operator", Turkish Journal of Electrical Engineering & Computer Sciences, vol. 12, no. 1 **(2004)**, pp. 43-52.
[26] K V S R P Varma, A Apparao, T Sita Mahalakshmi, P V Nageswara Rao, Narasimha Rao Kandula, "A Computational Intelligence Technique for Effective Diagnosis of Diabetes Disease using Genetic Algorithm", Proceedings of National Conference (NCETCS2014), **2014**, pp. 289-295.
[27] K William C., P H. Bennett, R F. Hamman, and M Miller., "Diabetes incidence and prevalence in Pima Indians: a 19-fold greater incidence than in Rochester, Minnesota", American Journal of Epidemiology, Vol. 108, No. 6 **(1978)**, pp. 497-505.
[28] S X-J and L, H, "A genetic algorithm-based approach for classification rule discovery", In Proceedings of the IEEE International Conference on Information management, Innovation Management and Industrial Engineering (ICIII08), **2008**, pp. 175-178.

