Analytical Study of Heart Disease Prediction Comparing With Different Algorithms

Sana Bharti
M.Tech Scholar, ASET-CSE
Amity University
Noida, India
bhartisana90@gmail.com

Dr.Shaliendra Narayan Singh
Associate Professor
Amity University
Noida, India
snsingh36@amity.edu

***Abstract*— In data mining there are several ways, approaches to predict any disease and different researches are still going on. In this survey, we have studied several algorithms (like genetic algorithm, Particle Swarm Optimization, Artificial Neural Network) which play very essential role in determining or predicting heart disease. Here we firstly describe the basic concepts of these three algorithms ,and analyze how these algorithms helps in prediction of heart diseases.**

***Keywords*—** *Heart disease; Particle Swarm Optimization;Genetic Algorithm; Artificial Neural Network*

I. INTRODUCTION

Data mining one of the technique of extracting the knowledgeable information from a large data set [1]. When combining with several machine learning algorithms, population based algorithms, machine learning can lead to better performance with high accuracy. Data mining extracts the clinical attributes and pathological data and generates biological hypothesis from the large medical data set[2].

Heart disease is the form of cardio-vascular disease which is prevailing in the whole world like a wave and becoming a bigger cause of deaths. As quoted by Mr. KK Aggarwal ,President of Heart Care Foundation of India "2.4 million Indians are dying due to heart disease every year. The number will continue to increase due to things like stress, unhealthy eating habits, lack of physical exercise, lack of sleep and dependence on alcohol and cigrattes".

Data mining is playing very essential role nowadays in the biomedical field in predicting various diseases. Now a days due to various changes in the environment, unhealthy eating habits person suffer from multiple disease of the same category so physicians can not able to predict right disease. In In this survey we have found that data mining concepts when use with intelligent algorithms can help in prediction of medical disease having multiple inputs and helping the doctors to tackle with such kind of problems.

II. ARCHITECTURE TO PREDICT HEART DISEASE

There has been numerous ways to predict the risk of heart disease, but there is a basic flow of predicting heart disease which is proposed by Bhuvaneswari Amma is described below in figure1 [11].

A. Database of patients

The database of patients is been given which have 14 major components which helps in predicting heart disease: age(in years),sex(m/f),chest pain type(typical angina, atypical angina, non angina , asymptomatic),resting blood pressure(in mm-Hg),cholesterol(mg/dl), high fasting blood sugar(1/0),resting electrocardiographic results(1/0), maximum heart rate achieved, resting heart rate , exercise induced angina(1/0), St depression induced by exercise relative to rest, number of color vessels, obesity , thal(3=normal, 6= fixed defect, 7=revesrsable defect).

B. Analysis of data

It is one of the most important steps as the data in the database contain most of the redundant and noisy data so by analysis of data, we can perform data cleaning, data integration fill up missing values, removal of redundant data because handling missing value and redundant data would lead to incorrect output.

C. Feature selection

It is processing step which helps in reducing the dimensionality and thus increasing the accuracy and performance. PCA, Chi square test are the some of the techniques for feature selection. PCA extract the useful information from the database and name them as principal components. Chi square test finds the worth of the attribute

D. Optimization Algorithm

Various different algorithms we can apply here to explore the best attribute which will participate in reproduction by evaluating the fitness value which is assigned to each attribute(individual). Algorithms which can be used can be PSO, Genetic Algorithm, Ant Colony Optimization, Artificial Neural Network or may be the hybrid of any of these algorithms.

ISBN:978-1-4799-8890-7/15/$31.00 ©2015 IEEE

E. Training and Classification

Input data are trained and several classification techniques are applied so that they extract hidden useful information and give more accurate results.

F. Prediction Engine

Predicts the whether the person has a heart disease or will suffer in future.

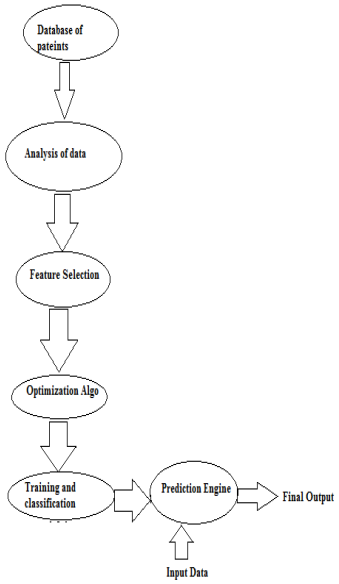

Fig. 1. Basic flow of prediction of heart disease [11]

III. BASIC CONCEPTS

A. Particle Swarm Optimization

Particle Swarm Optimization is invented by Kennedy and Eberhart ,which imitate the behavior of bird flocking[13]. Particle Swarm is a group of birds or fish schools in which all these particles are searching for food and attain a better position in each iteration. But they do not know how far the food is in each iteration. Then what will be the optimal strategy to find the food? The efficient way is to follow the bird which is nearer to food. In our paper we relate Heart Disease prediction with the particle swarm intelligence algorithm. .

PSO is initialized with a group of random particles in an N dimensional space and each particle has its own velocity and position which is being calculated by using these 2 equations denoted by V_i and then searches for optima by updating generations[13].

$pre^{k}_{0}=pre_{min} + \mu(pre_{min}- pre_{max})$ 1)

$V^{k}_{i}= pre_{min} + \mu(pre_{min}- pre_{max})$ 2)

In each iteration particle is updated with two values and record them. The first one is the best solution, it has achieved so far . The fitness value is stored in memory . This value is called P_{best} . Another "best" value that is tracked by the particle swarm optimizer is the best value, obtained so far by any particle in the population. This best value is a global best and called G_{best}.

After finding these two values we update the velocity and position of particle by using the below two equations.

$V_{i+1}^{k}[] = w*V_i^{k}[] + b1* \mu() *(P_{best}^{k} [] - pre_i^{k}[]) +b2*\mu()*(G_{best}^{k}[]-pre_i^{k}[])$ 3)

$pre^{k}_{i+1}[]=pre^{k}_{i}[]+V^{k}_{i+1}[]$ 4)

V_{i+1} is the velocity of kth particle at i+1 time , pre is the present velocity of particle, w is the inertia factor which controls the velocity of the particle by weighting the contribution of previous velocity and controlling the how much previous particle's flight direction will influence new velocity , its value range[0,1] to have good convergent behavior and normally its value taken as 0.7298, μ is a number between (0,1) ,it is used to have good coverage of the problem space and avoid entrapment in local minima ,b1,b2 are positive learning vectors and usually b1=b2=2. If the V_{i+1}^{k} exceed the given value then velocity at that position is equal to V_{max} . **The pseudo code of PSO** is as below.

Step 1 Take the training data.

Step 2 Initialize the particles population with their position and velocity parameters(by using equation 1 and 2) .

Step 3 Evaluation of individual particle by calculating the fitness value ,if fitness value > P_{best} then update current value as P_{best}.

Step 4 Select the particle which has best fitness value among all the particles.

Step 5 Calculate particle velocity and position according to equation 3 and 4.

Step 6 Do until either minimum error is not attained or up to maximum iterations.

1) Strengths and Weakness Of PSO.

a) Strengths of PSO [7]

- PSO has the capability to handle difficult problems.
- PSO has availability of solutions.
- By using Guarantee Convergence PSO (GCPSO), it guarantees convergence towards local minima.
- PSO is more robust and has a property of inherent parallelism .

b) Weakness of PSO

- Sometimes PSO fitness function and search techniques are not obvious [7].
- PSO is comptunially intensive [7].
- PSO deals with difficult parameter optimization [7].

B. Genetic Algorithm

Genetic Algorithm is invented by John Holland in 1970.GA works on the principle of genetics and evolution as seen in reproduction process of human population[13]. The Genetic Algorithm follow principal is "Survival of Fittest" while searching and generating the solutions (individuals) which are then adapted to their environment. GA hasa population of strings (chromosomes or genome) that encode a candidate solution(phenotype) which leads towards a better solution. Solutions are usually represented in 0 or 1. The

evolution starts from randomly generated individuals as seen in biological population .In every iteration (generation) fitness of every individual is calculated and based on their current fitness value individuals are selected and modified to form a new population. GA terminates either satisfactory level of fitness is reached or maximum generations have been produced. Requirement of GA is as follows

a) Genetic representation of solution domain
b) Fitness function to evaluate solution domain.

The Genetic Algorithm is helpful in solving complex design optimization problem as it can solve continuous and discrete variables and non linear function without gradient information [2].

1) *Algorithm of GA:*

GA takes the fitness value of each individual which is calculated in each iteration to form new population having fittest members. Flow of GA is shown in figure 2.

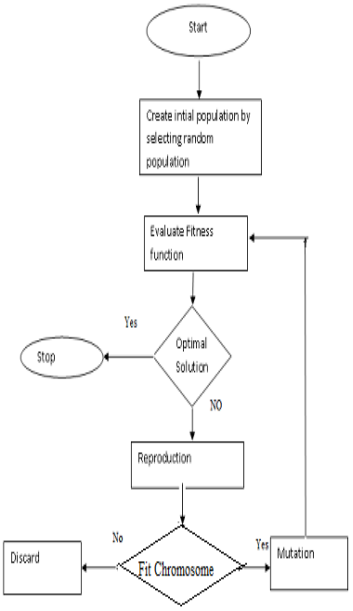

Fig. 2. Flow of Genetic Algorithm [13]

C. Neural Network

ANN is an another way of which helps in predicting heart disease ANN behavior is just like biological neural network, a dense collection of neurons. Neural network implements a problem parallel in each layer level. It consists of three layers a) Input layer b) one or more hidden layer c) an output layer. Number of hidden layers and the number of neurons are totally dependent upon the complexity of the system. Output of input layer will be the input of next layer (hidden layer) and so on. In detecting heart disease parameters of input will be the patient's risk factors. ANN is trained by using following three methods.

a) Supervised Training: Supervised training is a machine learning task of gaining knowledge from labeled training data, i.e an external person gives neural network input data and its actual output and seeks to build a predictive model that will able to generate correct predictions of new data.
b) Unsupervised Training: Unsupervised is also a machine learning task which predicts the output from dataset consisting of input data without labeled responses. It is an iterative learning process in which input is given to the neural network and weights are associated with inputs are adjusted or updated each time. Cluster analysis is one of the methods of unsupervised training
c) Reinforcement Training: Reinforcement Training is an automated machine learning task which allows the neural network to observe and learn its behavior based on feedback from the environment, this behavior can be learned once and keeps on updating it. Reinforcement training can converge to a global optimum solution which can be ideal behavior and maximize the performance.

1) Advantages of neural network:

- Due to the weights given by input makes neural network more robust
- Performance of neural network is improved by learning in each iteration.
- For solving the complex problem, a neural network can be parallelized to achieve better performance.
- Once all the training has been done, high accuracy with low error rate is being achieved.

D. PSO Vs Genetic Algorithm

PSO and GA both are evolutionary technique and both have some common features like a) both algorithms have a group of randomly generated population, b) both algorithms use fitness value to evaluate the solution and update the population accordingly, c) both the algorithms reproduce the new solution based on fitness value. Besides these common factors the way of traversing the search space is totally different [8].

Difference between PSO and GA:

- PSO does not have genetic operators like crossover, mutation. Particles update themselves according to the internal velocity of particles and have memory to store the G_{best} and P_{best} value due to which PSO is being used by many researchers.
- The mechanism of sharing information is also different, PSO is a one way information sharing mechanism as only best neighborhood give information to others and due to this all the particles converge to best solution which is mostly observed in local version only, whereas GA chromosomes share their information with each other so the entire population moves like group towards an optimal solution [9].

- Selection process in GA follows "survival of the fittest", there are several selection methods in GA like roulette wheel selection, etc. Apart from from selection in GA, an elitist strategy is used which results chromosome with highest fitness value is passed to the next generation. PSO does not follow any selection method, all the particles will continue to be the members of the population during the entire process [9].
- In GA crossover occurs between two randomly selected parents and the outcome of new chromosome is the mixture of the genetic material of two randomly selected parents. In PSO particles does not exchange their material with other particles instead particle influenced by its own previous best position and the best position in the neighborhood.
- Although PSO and GA approximately yield the same quality of solution but PSO is more computationally efficient and use less number of function operator.

IV. LITERATURE SURVEY

Many studies have been done and more and more is going on in prediction of heart disease to get more accurate results. They have applied several data mining techniques for analyzing and prediction of disease and achieved different results of accuracy by applying different methods.

- Heart Disease Prediction System proposed by AH Chen, SY Huang, PS Hong, CH Cheng, EJ Lin and is developed by using data mining technique a) selection of important attributes for prediction of heart disease b) artificial neural network for classifying heart disease based on important features . The accuracy of prediction is nearly 80% and develop a user friendly HDPS and has features like ROC curve display, performance display section [10].
- Measuring the performance level by using regression , confusion matrix, and estimate the ROC value before applying PSO feature reduction algorithm and after applying PSO feature reduction algorithm and in both the method artificial neural network is used for classification of patients as disease or non disease. Sivagowry.S and Dr. Durairaj.M has proved that after applying the PSO algorithm the performance level has been increased [14].
- Digaonsis of heart disease by M.Akhil Deekshatulu and Priti Chandra uses Artificial Neural Network and feature subset selection for classifying the heart disease .They have used PCA to reduce the number of attributes which reduces the irrelevant tests of patients in diagnosing the heart disease and has an accuracy level of prediction is 92.8% [1].
- Data mining model has been proposed by A. Sheik Abdullah for detecting Coronary Heart Disease and predicting the various events related to each patient record by using PSO with J48.This model reduces the number of attributes using feature selection. This model has selected 9 attributes and accuracy is 60.74%, which quite good as compared with other optimization algorithm like stepwise forward selection and stepwise backward selection [12].
- To solve the problem of premature convergence of C mean clustering based on PSO and which is less effective in handling the data set whose dimensions are larger than the samples, Qiang Niu has proposed a method , a novel fuzzy c-means clustering method based on enhanced PSO algorithm.It has 2 steps:
  a) Distribute the membership based on the distance between the sample and cluster center, making the membership meet the constraints of FCM.
  b) Optimization strategy is applied so that optimal particle can be guided to the close group.
- T.John Peter uses different classification models and detect the complex nonlinear relationships between dependent and independent variables in the medical data set and detect all possible interactions between predictor and variables. He has followed certain steps
  a) Change the data set into arff format.
  b) Apply attribute selection and various classification methods like naïve bayes, KNN , NN .
  c) Measure the accuracy, and found that that Naïve Bayes gives the better results in predicting the heart disease [3].
- Bhuvaneswari Amma N.G presented a medical diagnosis system which predicts the risk of cardiovascular disease by combining the advantages of neural network and genetic algorithm. Backpropagation Algorithm is used for training the data and due to this mean square error is being reduced. The classification accuracy of the training data set is 99% [11].
- Raina Hassan, Babak Cohanim, Olivier de Weck compares the performance of genetic algorithm(GA) with particle swarm intelligence(PSO) using a set of benchmark test problems and two space system design optimization problem and found that PSO outperforms the GA with a large differential in computational efficiency when solve unconstrained nonlinear continuous problem and less efficient when solve constrained nonlinear continuous problem [6].

V. CONCLUSION

The main objective is to study various algorithms that can be used to predict the heart disease and compare them to find out the best method of prediction of disease. We have found that on combing these algorithms with various data mining techniques like clustering, classification, association rule, etc or hybrid these algorithms with each other will lead to better performance and high accuracy rate .In my future work I will

use Particle Swarm Optimization using feature subset selection and classification data mining techniques to predict the heart disease.

ACKNOWLEDGMENT

I would like to thank Dr.Shaliendra Naryan Singh my mentor for sharing his pearls of wisdom with me during this survey. I am immensely grateful to my mentor for his comments on an earlier version of the manuscript, although any errors are my own and should not tarnish the reputation of the esteemed person.

.

REFERENCES

[1] M. Akhil Jabbar, B.L Deekshatulu & Priti Chandra, **"** Classification of Heart Disease using Artificial Neural Network and Feature Subset Selection", Global Journal of Computer Science and Technology Neural & Artificial Intelligence, Volume 13 Issue 3 Version 1.0 Year 2013

[2] Ms. Preeti Gupta ,Ms. Punam Bajaj, **"** Heart Disease Diagnosis System Based On Data Mining And Neural Network", international journal of engineering sciences & research technology, June, 2014.

[3] T.John Peter, K. Somasundaram, **"** an empirical study on prediction of heart disease using classification data mining techniques", IEEE-International Conference On Advances In Engineering, Science And Management (ICAESM -2012) March 30, 31, 2012.

[4] Qiang Niu, Xinjian Huang, **"** An Improved Fuzzy C-means Clustering Algorithm based on PSO", JOURNAL OF SOFTWARE, VOL. 6, NO. 5, MAY 2011.

[5] Sunita Sarkar,Arindam Roy,Bipul Shyam Purkayastha, **"** Application of Particle Swarm Optimization in Data Clustering: A Survey", International Journal of Computer Applications (0975 – 8887) Volume 65– No.25, March 2013.

[6] Rania Hassan Babak Cohanim Olivier de Weck, **"**a copmarison of particle swarm optimization and the genetic algorithm",46 AIAA Structue,Strctural Dynamics and and Material Confrence ,18-21 April 2005,Austin ,Texas

[7] Dr. G. M. Nasira, S. Radhimeenakshi," Particle Swarm Optimization- A Review on Algorithmic Rule, Application and Scope", International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064

[8] Azhar W. Hammad, Dr. Ban N. Thannoon, **"** genetic algorithm versus particle swarm optimization in n-queen problem", Journal of Al-Nahrain University Vol.10(2), December, 2007, pp.172-177.

[9] Russell C. Eberhart and Yuhui Shi, **"** Comparison between Genetic Algorithms and Particle Swarm Optimization", EP'98 Proceeding of the 7th International Conference on Evolutionary VII Pages 611-616.

[10] AH Chen, SY Huang, PS Hong, CH Cheng, EJ Lin, **"**HDPS: Heart Disease Prediction System"10 Department of Medical Informatics, Tzu Chi University, Hualien City, Taiwan.

[11] Bhuvaneswari Amma N.G**"**Cardiovascular Disease Prediction System using Genetic Algorithm and Neural Network",IEEE

[12] Sheik Abdullah, **"** A Data Mining Model to Predict and Analyze the Events Related to Coronary Heart Disease using Decision Trees with Particle Swarm Optimization for Feature Selection", International Journal of Computer Applications Volume 55– No.8, October 2012.

[13] Sapna Katiyar**,"** A Comparative Study of Genetic Algorithm and the Particle Swarm Optimization **".**

[14] Sivagowry.S , Dr. Durairaj.M, **"** PSO - An Intellectual Technique for Feature Reduction on Heart Malady Anticipation Data",International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 9, September 2014.