International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2015): 78.96 | Impact Factor (2015): 6.391

Genetic Optimization of Pattern Classification Using Heart Disease Data

Kamini Viswakarma[1], Satish Dehariya[2]

[1, 2]Samrat Ashok Technological Institute, Vidisha, Madhya Pradesh., India-464001

Abstract: ***Classification is one of the data mining techniques which is generally used in the prediction of data. Particularly in pattern classification a function is built to map input space into output space using multiple classes. For this purpose, artificial neural networks are very effective to build a model and to test and train the data. For the training process many training algorithms are used to increase the performance. In this paper, we shall make a training of feed forward neural network using Backpropagation algorithm, which is optimized with Genetic Algorithm with heart disease dataset with multiple class. Graphs are plotted to show the accuracy of the classification.***

Keywords: Data mining, classification, artificial neural network(ANN), Genetic algorithm (GA), Back-Propagation Algorithm(BPN)

1. Introduction

Data mining is the field where computer science and statistical field coincide. Here, we try to get any pattern or we can say that any high level information from a low level real world dataset. The information we get from the pattern is used in next stage of operations [1]. Data mining plays a vital role in KDD (Knowledge Discovery in Databases). Data get the chance to work because researchers want feasible information out of a large dataset in a affordable time. The most common operation in current data mining functions include Classification, Clustering, Discovering association rules, summarization and sequence analysis. Classification is one of the important techniques of data mining. The input for the classification problem is an input data-set that can be defined as the training dataset having a number of properties otherwise known as attributes. The attributes can be continuous or categorical. The categorical attributes can be a class label or the classifying attribute. The primary aim is to use the training dataset to build a classification model or a set of rules for the classification operation which is class label based on the properties of other class attributes such that the model can be trained according to the current data and will be used to classify new data recorded from same source and not from previous input training dataset [5].

Data mining problems can be effectively solved by many soft computing techniques like fuzzy theory, neural networks, genetic algorithm etc which can give us smart, effective, manageable, economical solution than the previous methods. Hence, from above techniques we take Artificial Neural Network (ANN) for the study of classification of the dataset. Here, we use both sequential and parallel processing technique for the dataset [2], [8]. Data mining techniques are widely used in health care applications like analysis of clinical data of any field of medical science to get a better classification of the data from the patients. The resulting model shows a common information or we can say that a formalized knowledge, which can be provide us with a better model for obtaining a better diagnostic opinion.

1.1 Artificial Neural Networks

Artificial Neural Networks(ANN) are networks which are normally inspired by biological neural networks consist of neurons. Here, computer scientists try to copy the function of human brain into computer networks [4, 24 & 26]. ANN are used in many application areas like pattern recognition and classification. The ability of ANN to make decision is more useful, when mainly based on sum total of all input patterns where in traditional methods we have to take each input pattern into consideration to get a optimal conclusion. ANN is pretty much compatible with parallel processing a well as serial processing at each level. These networks come with another significant characteristic i.e. fault tolerance [10], [12] [19]. They can be categorized into two types based on the training method of the classification model: Supervised training and unsupervised training. Networks that are called supervised need the actual desired output neuron for each input neuron whereas unsupervised networks do not need the desired output neuron for each input neuron.

Unsupervised learning: In this learning method, the targeted output is not given to the network which is much similar to a case where a teacher is not present while a student learns. [22] The network learns from the successive input pattern and get adaptive to it. Mainly, the network finds an inter-relation between the input patterns.

Supervised learning: - It works like an external teacher is present to teach so that each output neuron is instructed and compared with its desired output neuron and corresponding input signals. The input pattern is processed and the output is calculated [21]. If the output does not match with the targeted output, then the corresponding weight is adjusted to get closer value towards the output.

Then, we discuss about various neural network techniques to process the data sets we have.

1.2 Back Propagation Algorithm

The back propagation algorithm is commonly used as a learning algorithm which learns from its error while

discovering the classification rule. Back propagation is a form of supervised learning for multi-layer feed forward neural networks [3]. It is first proposed by McCulloch and Pitts [27]. It is very easy to implement. Here, one input layer(x1,x2,…,xn) is taken according to the requirement. At least one hidden layer and one output layer(y1,y2,…,yl) as shown in Fig.1

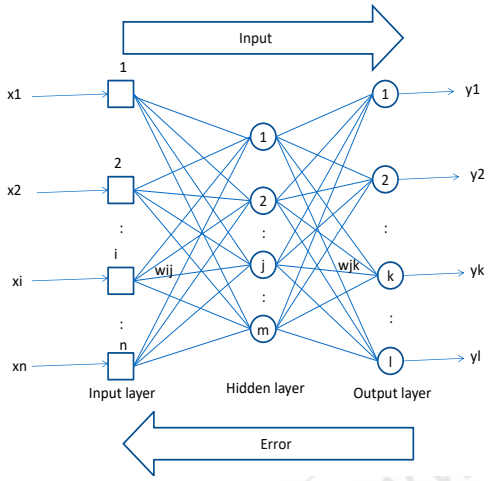

Figure 1: ANN with backpropagation algorithm

Firstly, the input signal is given through the input layer and then it is propagated towards the output layer. After the error is calculated that error signal is propagated towards the back means towards the input layer for which it is known as backpropagation algorithm [18].

1.3 Algorithm

Step 1: Initialisation
All weights and the bias for each neuron is initialised. The values of these randomly initialised weights are kept within a small range which is suitable for the activation function used in the classification model. The bias is always in the negative [20].

Step 2: Activation
Application of inputs to the neural network for each input neuron with input data like $x_1(p)$, $x_2(p)$,…, $x_n(p)$ and desired outputs to output matrix as $y_{d,1}(p)$, $y_{d,2}(p)$,…, $y_{d,n}(p)$ to activate the network.

(a) Calculation of output for the hidden layer using the sigmoidal activation function:

$$y_j(p) = sigmoid\left[\sum_{i=1}^{n} x_i(p)\cdot w_{ij}(p) - \theta_j\right] \quad (1)$$

where n is the total number of inputs for total no. of input neurons j in the hidden layer, and *sigmoid* represents the *sigmoidal* activation function of the classification model.

(b) Calculation of input for the output neuron of the model is done as the same way calculation for the hidden layer.

$$y_k(p) = sigmoid\left[\sum_{j=1}^{m} x_{jk}(p)\cdot w_{jk}(p) - \theta_k\right] \quad (2)$$

where m is the total number of inputs of neuron k present in the output layer of the model.

Step 3: Weight training
Updating of weights are done by below formula that is associated with back propagating error calculated. Calculation of the error gradient for the total neurons available in the output layer is done as follows:

$$\delta_k(p) = y_k(p)\cdot[1-y_k(p)]\cdot e_k(p) \quad (3)$$

Where $e_k(p) = y_{d,k}(p) - y_k(p)$

Calculation of correction for the weights:

$$\Delta w_{jk}(p) = \alpha \cdot y_j(p)\cdot \delta_k(p) \quad (4)$$

Change in weights at the level of all available output neurons for next iteration:

$$w_{jk}(p+1) = w_{jk}(p) + \Delta w_{jk}(p) \quad (5)$$

(*b*) Calculation of the error gradient for the all available neurons in the predefined hidden layer of the model:

$$\delta_j(p) = y_j(p)\cdot[1-y_j(p)]\cdot\sum_{k=1}^{l}\delta_k(p)\, w_{jk}(p) \quad (6)$$

Calculation of weight correction in the level of hidden layer of the model:

$$\Delta w_{ij}(p) = \alpha \cdot x_i(p)\cdot \delta_j(p) \quad (7)$$

Change in weights for the hidden layer neurons for next iteration:

$$w_{ij}(p+1) = w_{ij}(p) + \Delta w_{ij}(p) \quad (8)$$

Step 4: Iteration
Increase iteration *p* by one, go back to *Step 2* and repeat the process until the selected error criterion is satisfied.

1.4 Advantages

It is easy to implement. It is simple. It can be implemented efficiently on the parallel architecture and parallel processing of computer.

1.5 Disadvantages

Backpropagation network is slower to train than other types of networks. It may stick in a local minimum for which it can't calculate the global minima. The major drawback is it is not adaptive to the input pattern of a dataset.

2. Related Works

- In 2003, Back Propagation algorithm based Neural Network is used for Classification of IRS-1D Satellite Images by E. Hosseini Aria, J. Amini, M.R.Saradjian[20]. In this classification process one 3 layer network used and total 6 classes are taken. In the same year, N Kannathal, U Rajendra Acharya, Choo Min Lim, PK Sadasivan & SM Krishnan used Ann and SOM classifier for analysis of ECG signals using cardiac dataset.[21].
- In 2007, Insung Jung, and Gi-Nam Wang proposed a dividend exclusive connection based BPN for pattern classification. For the simulation process random data was taken. The performance of the network was little better than that of standard BP[18].
- In 2012, a classification model based on BPN for parameters for Tsunami was given by M. Umadevi and S.

Srinivasulu. Here, standard BP algorithm is used for calculation [3].

- In 2011, Dr. K. Usha Rani given an parallelism procedure to get analysis on heart disease dataset using neural network [10].

3. Proposed Work

We can use genetic algorithm to optimize pattern classification using current artificial neural network (ANN) to get a better classification [25].

3.1 Genetic Algorithm

Genetic algorithm plays a vital role in evolutionary computing. GA are normally an implementation of Darwin's theory about evolution in biological reproduction which is popularly known as "survival of the fittest" [6], [7]. Genetic algorithms are the methods of solving real life problems by applying and copying the ways that nature adopts while transmitting characters from parents to Childs like Selection, Crossover, Mutation and Accepting to get a solution to that problem which can be called as termination criteria [9], [11].

3.1.1 Selection

After the formation of chromosome, a population is formed in a desired range. All the chromosomes in this population go through a fitness function which is designed according to the mathematical model to generate the cost for all the chromosomes. Then according to the cost of chromosomes, the fittest among them is chosen as parent for the reproduction process [13].

3.1.2 Reproduction

The chosen parents from the selection process are used to crossover to generate a whole new population. The crossover process can be of different types like single point crossover, two-point crossover, and uniform crossover [14].

3.1.3 Crossover

Crossover [15] is a genetic operator in which two chromosomes mate to give birth to a new one. This is the idea to get a new chromosome so that the new offspring may get the good characteristic from the both the parents and may be better than that of the both of the parents. There are several types of crossover like one-point crossover, two-point crossover, uniform crossover, arithmetic crossover and heuristic crossover. In the simulation process single point crossover has been taken into account.

3.1.4 Mutation

Out of the new population generated from the reproduction process, some of the child chromosomes are chosen randomly over the population for mutation. The no. child chromosome chosen for mutation depends upon the mutation rate. [16] The mutation can be of different types like fixed point or random point mutation. In this simulation uniform mutation is adopted. After the mutation process, the entire population again goes for selection process. Thus, generation after generation is produced until the optimal solution is found.

3.1.5 Termination

The termination process is always user defined as the termination of the algorithm always depends upon the methodology [17]. There are some stopping criteria which are normally used are as follows .

1) Maximum no. of generation reached.
2) Optimal solution is found.

3.1.6 Steps of genetical algorithm

Step:1
Represent the collection of Weights and bias as Genes Of the chromosome. Collection of these genes are represented as a Chromosome.

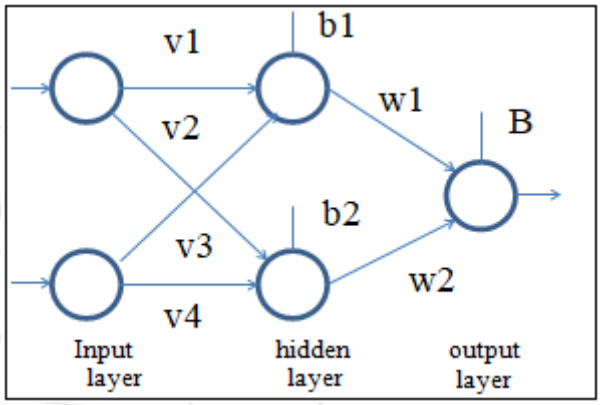

Figure 3: ANN with weights and bias

Chromosome = [v1v2v3v4w1w2b1b2B]
Step : 2
Initialize the chromosomes randomly.
Step : 3
Feed Forward the inputs towards the outputs.
Step:4
Calculate the Fitness Of chromosomes.
Step:5
Sort the Weights according to their Fitness.
Step:6
Replace all the lowest fitness chromosome with highest fitness chromosome.
Step:7
Perform Cross-Over.
Step:8
Find the termination Condition and
Repeat Step-2 to Step-7 until it is satisfied.

Pseudo code for genetic algorithm:

```
S=no. of solution, C=no. of chromosome,
 I=no. of Iteration, G=no. of genes
Initialization
        generate S solutions randomly;//named as P
        Select best solution from S and save as P1;
Error Calculation
For i=1 to C
   P[i, G+1]=error; //error calculated from the feed forward
network
 endfor
Loop for condition
        for i = 1 to I do
                total no. of P=C*G;
                Pop=sortrows(Pop);
                Crossover(Uniform)
                for j = 1 to C do
                        randomly select two solutions
                X1 and X2 from P;
                        generate X3 and X4 by Uniform
                crossover to X1 and X2;
                        save X3 and X4 to P1;
                endfor
                Updating
                        update P=P1+P2;
        endfor
Returning the best solution
return the best solution X in P;
```

3.2 Advantages

The order of training procedure is much greater than that of Backpropagation algorithm. It has a clean parallel structure which is quite easy for calculation in comparison to other neural networks [4]. As many nos. of training set increases, the network converges to an optimal classification.

3.3 Disadvantages

Not as general as backpropagation. This require a large memory. It requires a representative training set for each test dataset for classification.

4. Result and Discussion

From the backpropagation algorithm, a program is developed to classify heart disease dataset and another program is developed to optimize the classification using genetic algorithm. The Heart dataset has been collected from UCI Machine Learning Repository. It is very simple which has only 2 classes. It have 13 attributes like Age real [29, 77], Sex binary [0, 1], ChestPainType nominal [1, 4], Resting Blood Pressure real [94, 200], Serum Cholestoral real [126, 564], Fasting Blood Sugar binary [0, 1], Resting Electrocardiographic nominal [0, 2], Max Heart Rate real [71, 202], Exercise Induced binary [0, 1], Oldpeak real [0.0, 62.0], Slope ordered [1, 3], MajorVessels real [0, 3], Thal nominal [3, 7], Class {1, 2}. A total of 270 observation has been used for this simulation.

The simulation process is carried on a computer having Dual Core processor with speed 2.13 GHz and 4 GB of RAM. The MATLAB version used is R2015a.

After all the training process carried out using both back propagation algorithm and optimized classification method using genetic algorithm, the simulation results has been recorded. All this results has shown in Table-1 and the graphs are plotted to show the simulation in fig.4 and fig.5.

Table 1: Comparison of classification between GA and BPN algorithm

Algorithm	Genetic Optimization	BPN Algorithm
No. of Epochs	1000	100000
RMSE	0.0161	0.0201
Classification Accuracy (In %)	92.9630	84.0741
TOC (Time of Calculation)	1769.853396	3920.397029

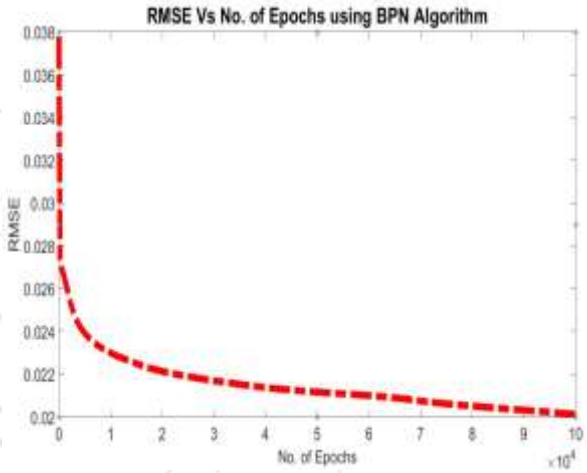

Figure 4: RMSE Vs. No. of Epochs using BPN Algorithm

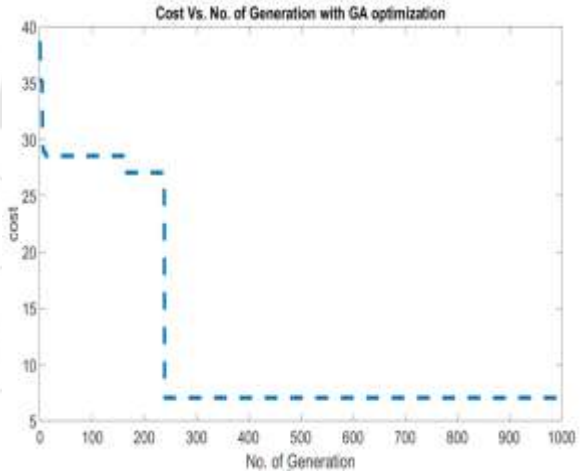

Figure 5: Cost Vs. No. of Generation with GA optimization

From the graph and the table we can see that using backpropagation algorithm, we have successfully classified heart dease data set to an extent of 84.07% accuracy with an time of 3920.39 seconds, which is quite long time and less accuracy. .For that genetical algorithm is adopted to have a higher accuracy in comparison to back propagation algorithm and relatively low time for the output layer (Table-1) in comparison to backpropagation. The classification accuracy is around 93% for heart disease dataset.

5. Conclusion and Future Work

From the above simulation process, we can consider GA optimized ANN is a good classifier in comparison to ANN based on backpropagation algorithm. Besides, for a good classification a large no. of input dataset needed from real life. By a good hybrid network like PNN (probabilistic neural network), RBFN(radial basis function network) we can classify more data on various fields which will save time and will be a good tool for all data mining process like prediction, optimization etc.

References

[1] Dayanand Savakar, "Identification and Classification of Bulk Fruits Images using Artificial Neural Networks", International Journal of Engineering and Innovative Technology (IJEIT) Volume 1, Issue 3, March 2012
[2] Le Hoang Thai, Tran Son Hai & Nguyen Thanh Thuy, "Image Classification using Support Vector Machine and Artificial Neural Network", I.J. Information Technology and Computer Science, 2012
[3] M. Umadevi & S. Srinivasulu, "Parameters for Tsunami Classification Model using BPN Based Artificial Neural Network", European Journal of Scientific Research, ISSN 1450-216X Vol.76 No.4, pp.633-647, 2012
[4] S. Daniel Madan Raja & A. Shanmugam, "ANN and SVM Based War Scene Classification Using Invariant Moments and GLCM Features: A Comparative Study", International Journal of Machine Learning and Computing, Vol. 2, No. 6, December 2012
[5] S. R. Patil & S. R. Suralkar, "Fingerprint Classification using Artificial Neural Network", International Journal of Emerging Technology and Advanced Engineering, ISSN 2250-2459, Volume 2, Issue 10, September 2012
[6] Singhai Rakesh & Singhai Jyoti, "Registration of Satellite Imagery Using Genetic Algorithm", Proceedings of the World Congress on Engineering 2012 Vol II WCE, July 4 - 6, 2012
[7] Valsecchi Andrea & Damas Sergio, "An Image Registration Approach using GeneticAlgorithms", WCCI 2012 IEEE World Congress on Computational Intelligence, 10-15, 2012
[8] Haoyu Ma, Bin Hu, Mike Jackson, Jingzhi Yan & Wen Zhao, "A Hybrid Classification Method using Artificial Neural Network Based Decision Tree for Automatic Sleep Scoring", World Academy of Science, Engineering and Technology, 2011
[9] Junli Lu, Guang Zhao, Cheng Yang & Junjia Lu, "New Fuzzy k-NN Classification by Using Genetic Algorithm", Seventh International Conference on Natural Computation, 2011
[10] K. Usha Rani, "Analysis Of Heart Diseases Dataset Using Neural Network Approach", International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.1, No.5, September 2011
[11] Selma Ayşe Özel, "A Genetic Algorithm Based Optimal Feature Selection for Web Page Classification", IEEE, 2011
[12] Steve Diersena, En-Jui Leeb, Diana Spearsc, Po Chenb & Liqiang Wanga, "Classification of Seismic Windows Using Artificial Neural Networks", Science Direct, 2011
[13] RC Chakraborty, "Fundamentals of Genetic Algorithms", AI Course Lecture 39-40, 2010
[14] Zhang Lianmei & Jiang Xingjun, "Research on New Data Mining Method Based on Hybrid Genetic Algorithm", International Symposium on Computer, Communication, Control and Automation, 2010
[15] Ingole V. T., Deshmukh C. N., Joshi Anjali & Shete Deepak, "Medical Image Registration Using Genetic Algorithm", Second International Conference on Emerging Trends in Engineering and Technology, ICETET, 2009
[16] Ji Peirong, Hu Xinyu & Zhao Qing, "A New Genetic Algorithm for Optimization", IEEE, 2008
[17] Xian-Jun Shi & Hong Lei, "A Genetic algorithm-Based Approach for Classification Rule Discovery", International Conference on Information Management, Innovation Management and Industrial Engineering, 2008
[18] Insung Jung & Gi-Nam Wang, "Pattern Classification of Back-Propagation Algorithm Using Exclusive Connecting Network", World Academy of Science, Engineering and Technology, 2007
[19] Abdulhamit Subasia & Ergun Ercelebi, "Classification of EEG signals using neural network and logistic regression", Computer Methods and Programs in Biomedicine, 2005
[20] E. Hosseini Aria, J. Amini & M.R.Saradjian, "Back Propagation Neural Network for Classification of IRS-1D Satellite Images", 2003
[21] N Kannathal, U Rajendra Acharya, Choo Min Lim, PK Sadasivan & SM Krishnan, "Classification of cardiac patient states using artificial neural networks", Exp Clin Cardiol Vol 8 No 4 2003
[22] Guoqiang Peter Zhang, "Neural Networks for Classification: A Survey", IEEE Transactions On Systems, Man, And Cybernetics—Part C: Applications And Reviews, Vol. 30, No. 4, November 2000
[23] Howard E. Michel & A. A. S. Awwal, "Enhanced Artificial Neural Networks Using Complex Numbers", 2000
[24] Gisbert Schneider & Paul Wredeb, "Artifcial neural networks for computer-based molecular design", Progress in Biophysics & Molecular Biology, 70,175-222, 1998
[25] Philipp Koehn, "Combining Genetic Algorithms and Neural Networks: The Encoding Problem", December 1994
[26] A. Pomerleau, Jay Gowdy & Charles E. Thorpe, "Combining Artificial Neural Networks and Symbolic Processing for Autonomous Robot Guidance", Engng App/ic. ArliJ. Inrell. Vol. 4. No. 4, pp, 279-285, 1991
[27] W. McCulloch and W. Pitts, "A Logical Calculus of the IdeasImmanent in Nervous Activity," Bulletin of Mathematical Biophysics,vol.5, pp. 115-133, 1943.

Author Profile

Kamini Vishwakarma received the B.E. degree in Computer Science and Engineering from Technocrats Institute of Technology, Bhopal in 2015 and pursuing M.Tech. degree in Computer Science and Engineering from Samrat Ashok Technological Institute, Vidisha. Implementation of E-business pattern using object oriented approach has been her major project topic during her bachelor degree. Optimization using genetic algorithm has been the broad research area in her master's.

Satish Dehariya received the B.E. degree in Information Technology from Samrat Ashok Technological Institute, Vidisha in 2007 and M.Tech. degree in Computer Science and Engineering from Barkatullah University Institute of Technology, Bhopal. He has been working in Samrat Ashok Technological Institute, Vidisha as teacher and Cloud computing has been the research topic since his master's. He holds ten years of experience in the various fields of data mining techniques.