RESEARCH ARTICLE OPEN ACCESS

Data Mining Using Neural–Genetic Approach: A Review

Parvez Rahi[1], Bhumika Gupta[2], Sombir Singh Bisht[3]
([1, 2,3]Dept. of Computer Science & Engineering, G. B. Pant Engineering College Pauri Garhwal, Uttarakhand)

ABSTRACT
In the advance age of technology, there is an increasing availability of digital documents in various languages in various fields. Data mining is gaining popularity in field of knowledge discovery. Data mining is the knowledge discovery process by which we can analyze the large amounts of data from various data repositories and summarizing it into information useful to us. Due to its importance of extracting information/ knowledge from the large data repositories, data mining has become an essential part of human life in various fields. Data mining has a very wide area of applications, and these applications have enriched the human life in various fields including scientific, medical, business, education etc. Here in this paper we will discuss the emphasis of Neural Network and Genetic Algorithm in the field of data mining.
Keywords – Data Mining, KDD, Neural Network (NN), Genetic Algorithm (GA), Back-propagation (BP).

I. INTRODUCTION

1. Overview of Data Mining

Data Mining is a process designed to analyze and explore the data in search of consistent patterns or to analyze the systematic relationships between data or variables, and then to validate the findings by applying the detected patterns to new subsets of data [1]. A formal definition of data mining is given as follows: Data mining is the non trivial extraction of implicit previously unknown and potentially useful information about data [2]. Data mining is an essential step in the knowledge discovery process that produces useful patterns from data. The terms of KDD and data mining are different. KDD maps whole data to discover useful knowledge from data. Data mining refers to discover new patterns from data in databases by applying intelligent methods and algorithms to extract useful knowledge [3].

Based on "Fig. 1", KDD process consists of iterative sequence methods as follows.

1. Selection: Data relevant to the analysis task is decided and retrieved from the database.

2. Pre-processing: Multiple data sources are combined to remove noise and inconsistent data

3. Transformation: Selected data is transformed in to forms appropriate for mining procedure.

4. Data mining: In this phase intelligent techniques are applied to extract patterns potentially useful.

5. Interpretation/Evaluation: In this phase redundant or irrelevant patterns are removed to interpret the pattern into knowledge; Translating the useful patterns into human understandable forms.

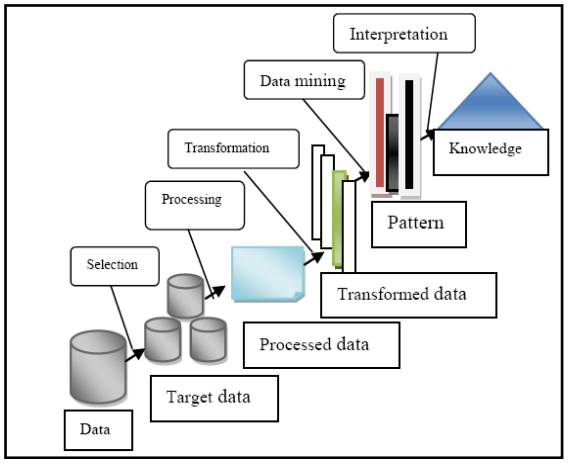

Fig. 1 KDD Process

2. Overview of Neural Network

A Neural Network (NN) is a collection of many Processing Elements (PEs), called "neurons" and all neurons interconnected to other neurons and each interconnection have a weight associated with it. Each PE performs many simple computations, like calculating a weighted sum of its input connections, and computes the corresponding output signal that is sent to other PEs as an input signal. The training (mining) of a NN is done by adjusting the weights (real valued numbers) of the interconnections, so that the NN may produce the desired output [6].

The Artificial Neural Network (ANN) is a very commonly used technique to solve data mining problems. Neural Network is a set of processing units which are assembled in a tightly interconnected network, based on some features of the biological neural network. As biological neural network or human brain learns by its surrounding, ANN learns by its past experience. The structure of neural

network provides an opportunity to the user to implement parallel concept at each layer level. A very significant characteristic of ANN is that they are fault tolerant in nature. ANNs are very good in situations where information is uncertain and noisy.
ANN are an information processing methodology that differs drastically from conventional methodologies in that it employ training by examples to solve problem rather than a fixed algorithm [4,5]. Training of a NN can be categorized in to two methods: Unsupervised training and Supervised training. Supervised networks that are require the actual desired output for each input and require a teacher for training, where as unsupervised networks does not require the desired output for each input and does not require a teacher for training.

Learning process of neural network is an iterative learning process in which every data cases are presented to the network one by one, and the weights adjustment is done for all the input values coming to network [6]. When all the cases are presented, the process starts again from its beginning. In the learning phase, weight adjustment is done to make the network learn so that it may able to predict the correct class label of input samples whose class label is unknown. Once the structure of network is ready for a specific problem, than the network is ready to be trained.

Initial weights are chosen randomly to start the training process. Then the learning or training, begins. The most popular and commonly used neural network training algorithm is back-propagation algorithm. Although many types of neural networks are available for classification purposes [7].

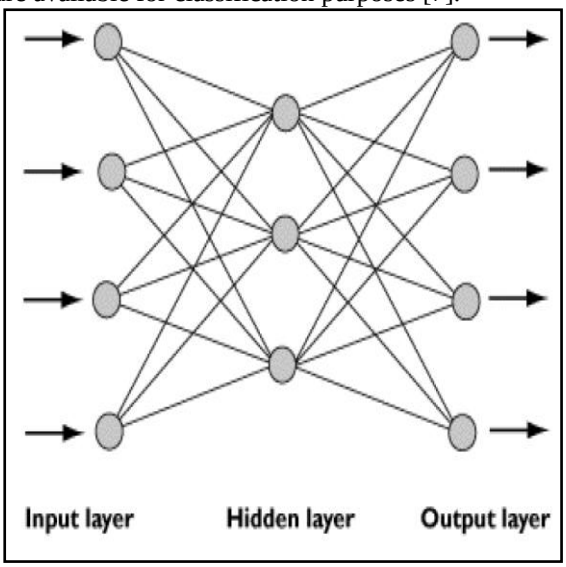

Fig. 2 Artificial Neural Network

3. Overview of Genetic Algorithm

Genetic algorithm is an adaptive heuristic random global and direct search method based on imitate of nature biological evolution mechanism, its essence is an efficient, parallel, global search method made up of replication, selection, crossover, mutation operator. Its implementation process includes encoding; create population; fitness calculation; replication; crossover; mutation and so on. The basic idea of genetic algorithms is based on evolution theory developed by Darwin and inheritance theory developed by Mendel. The most important of Darwin evolution theory is survival of fittest theory; it believed that all species can much more adaptable of environment with development. Basic characteristics of each species individual can be inherited by descendants, but future will have some new changes which different from parents. And only those characteristics which adapt to the environment can be preserved when the environment changes. The most important of Mendel inheritance theory is gene genetic; it is considered as genetic exist with genetic code in the cell and shown as genes in the chromosomes. Each gene has a special position and control a special characteristic of the individual, and individual with a special gene has a certain adaptability of the environment. Gene mutations and genetic hybrids can produce offspring with more adaptable to environment. Gene structure with higher adaptability can be preserved after natural selection procedure which to select the best and to eliminate the worst. The genetic algorithm encodes the solutions of the problems with "chromosome" and implement it with encode string. A group of "chromosome" which is assumption solution must be given before implementation of genetic algorithms. First, the assumption solution must be suited in the 'context' of the problem, and be selected follow the principle of survival of the fittest so that the much more adaptive "chromosome" can be chosen to replicate, and then a group of next-generation 'chromosome' can be produced through crossover and mutation. In this way, to evolve from generation to generation, and finally it can converge to the 'chromosome' which is the most adaptive to the context of the problem, and it is the optimal solution of the problem.

Genetic algorithm has three basic operating to groups:

3.1 Selection

Selection purpose is to select the excellent individuals from current population, so they will have the opportunity to act as parent to propagate descendants of next generation. Fitness value is the basis to select the individuals much more adaptive to the environment. The GA embodies the principles of Darwin's theory of survival of the fittest by selection, and adaptable individuals have higher rate to contribute descendant for next-generation. The

common used operators are roulette selection, random competition selection, best reserved selection, stochastic sampling with replacement selection, etc.

3.2 Crossover

Crossover is the most important genetic manipulation of the genetic algorithm. Crossover algorithm will exchange genes from two different individuals selected at the same location, and then a new entity created. By crossover operation, the individual of new generation wound combines the individual characteristics of two parents. Crossover operation contains the idea of the information exchange. The commonly used crossover operators are single-point crossover, two-point crossover, arithmetical crossover and uniform crossover.

3.3 Mutation

First, an individual should be selected randomly from the group and then change the value of a gene string structure data from selected individual with small probability, which is to change one or a number of values to the other allele with a certain probability for all individuals in the group. Mutation provides an opportunity to create new individual for the new generation. The common used mutation operations are simple mutation, uniform mutation, non-uniform mutation and Gaussian mutation, etc [9].

II. LITERATURE SURVEY

In this section, a brief about all the research papers reviewed and studied is documented. Also a brief about the work done in the same is included here.

Singh and Chauhan [10], Said that, In this advance age of technology companies gather more and more data about market trends and customers every year but to make data useful for business application it is necessary to model complex relationships between inputs and outputs or to find patterns in data. Neural networks are non-linear statistical data modelling tools generally used for data mining purpose. Neural Networks are widely used for data mining purpose due to their black box nature, even though they have shown their importance in many situations. Neural networks basically consist of three pieces: the architecture or model; and the activation functions; the learning algorithm. To store, recognize, and associatively retrieve patterns; to filter noise from measurement data; to solve combinatorial optimization problems; to control ill-defined problems neural networks are trained very efficiently. Neural networks can mine specific information from a mass of history information and that can be efficiently used in financial areas, so due to this applications of neural networks, financial institution use it for business forecasting since last few years. Author says that neural networks performed better than conventional statistical approaches in financial forecasting and prove them excellent data mining tool. So, the use of neural networks in data mining is a promising field of research especially given the ready availability of large mass of data sets and the reported ability of neural networks to detect and assimilate relationships between a large numbers of variables.

K. Usha Rani [11], According to author classification is one of the important techniques of Data mining. Classification approach is used to solve many real world problems in various fields such as science, business, industry and medicine. They also have been applied to classify various areas of medicine, like diagnostic aides, medicine, biochemical analysis, drug development and image analysis etc**.** Neural Networks is a very popular tool used for classification. Neural Networks does the classification task very efficiently. Here in this study neural network approach is being used to analyze Heart diseases dataset. Practically neural networks are known to produce highly accurate results in various applications. Artificial neural networks are generally used in a broad range of medical applications to help doctors to analyze complex clinical data which is very useful for further decision making. In this experiment the feed forward neural network model and back-propagation learning algorithm with momentum and variable learning rate is used to trained the Heart Diseases database. Various test data are given as input to the network to analyze the performance of the network. To speed up the learning process parallelism mechanism is implemented at each neuron in all hidden and output layers. Experimental results of this research show that neural networks technique provides satisfactory results for the classification task.

Setiono and Liu [12], According to author classification is one of the notable problem of data mining problems, getting great attention recently in the database community. Here in this paper neural network is used to presents an approach to discover symbolic classification rules. It is noted that neural networks have not been suited for data mining because how the classifications were made is not explicitly stated as symbolic rules that are suitable for verification or interpretation by humans. Here in this proposed approach, neural network is used to extract concise symbolic rules with high accuracy. The network is first trained in such a way that it can achieve the required accuracy rate. Network pruning algorithm is used to remove the redundant

connections of the network. The network is analyzed to find out the activation values of the hidden units, and in further step result of this analysis is used to generate classification rules. The experimental results clearly demonstrate the effectiveness of the proposed approach on a set of standard data mining test problems.

Ayodele et al [13], Said that email has become one of the fastest and most efficient forms of communication in modern society. The increase of email users can cause of problem like, email congestion, high volume of email messages could lead to un-structured mail boxes, email overload, unprioritised email messages etc. So to avoid all these types of problem in email communication it is necessary to classify the emails in various categories according to their importance. In this paper back propagation technique with multi-layer neural network is used for new email classification. Back propagation is a special kind of network that can be trained in such a way that it may recognize different patterns including text, images, and signals. Neural network (NN) is a tool which has the ability to learn by example with back propagation techniques. We generate accurate email categories and analyze the characters of emails and study the email conversation structure, the inputs of the NN are the important words in email messages. In this research it is shown that neural networks using back propagation technique can be successfully used for semi-automated email classification. Result of this research shows 98% success in email category classification and new neural network algorithm is compared with human participants the algorithm's performance seems to work well and better than the existing approach.

Xianjun Ni [14], Author said that the application era of neural networks in the data mining has become wider day by day. The efficiency of data mining methods can greatly improve by using the combination of data mining method and neural network model. It also will receive more and more attention. Neural network is a parallel processing network which generated on behalf of human brain as human brain learn by its surrounding as well as neural network learns by its past training experience. Basically it uses the idea of non-linear mapping, the structure of the neural network itself to express the associated knowledge of input and output and the method of parallel processing. Due to defects of poor interpretability, complex structure and long training time the application of the neural network in data mining was not optimistic. BP network is the neural network most commonly used in data mining. Problem encountered in BP network is that the training is practically slow, it is difficult to determine training parameters and it may fall into local minimum. To reduce these problems people generally adopt the hybrid method of combining artificial neural networks and genetic algorithms and get better results. On the other hand neural network has many advantages like distributed information storage, information, reasoning, parallel processing, and self-organization learning, and also has the capability of rapid fitting the non-linear data, thus we can see that it can solve many problems which are difficult for other methods.

Li et al [15], Said that data mining is rapidly growing era of research in several disciplines, like databases, pattern recognition, statistics and parallel computing and high performance etc. With this emerging trend of data mining, image data mining from data base is new and hot researching area. Although general data mining is just extracting knowledge from large amount of data. Here in this paper, author propose a new scheme for mining, called ARMAGA (Association rules mining Algorithm based on a novel Genetic Algorithm), to extract the association rules from an image database, Here ARMAGA representation is used to represent every image. In this algorithm, first take advantage of the genetic algorithm which is specifically designed for discovering association rules, and then we compare propose Algorithm and existing algorithm. The ARMAGA algorithm is more efficient in terms of the execution time due to avoidance of generating impossible candidates. And finally we compare the results of the ARMAGA with the results of GA and ARMA, Result of this hybrid approach is better than GA and ARMA through the experimental results and the theoretic analysis.

Dou et al [16], In this research, author proposed an efficient data mining technique for making quick response to users and providing a friendly interface to overcome the low efficient problem and provide users with real demanded rules. In this research author introduce a quick response data mining system (QRDM). The proposed system consists of two major sections. In first section GA is used to charge of mining maximal frequent item sets and show them to users. In second section association rules are deduce in terms of maximal frequent item sets and then scans the database for obtaining real support and confidence of those rules. If apriori algorithm is used to mine all frequent item sets in those data, then the candidate item sets of data will become very huge and it become a burden to scan database many times. This proposed method avoids mining rules through huge candidate item sets, it just mines maximal frequent item sets and scans those frequent item sets

in the database in which users are interested. Finally the system, scan the database for the real support and confidence and show them to users. Therefore, the proposed method not only save time in scanning the database repeatedly and also make quick response to the users, and also provide a friendly interface to users so that they can select their interesting rules to mine.

Srinivasa et al [17], In this research, author present a new network intrusion detection system based on genetic algorithm named IGIDS. To detect the intrusion a set of technique is used in which the information from known types of attacks to detect suspicious activities at network and host levels is collected. There are mainly two techniques which are used to detect intrusion named anomaly based and signature based. Here the proposed system is payload-based. The proposed system uses the destination address and service port numbers to build a profile for each port monitored, and it does not consider other header features. In this process data mining is apply on the DARPA dataset to get the important information and convert it into rules. New rules are produced by using these rules and their fitness is calculated using the evaluation function. It is decided on the basis of fitness of rules whether the new rules are intrusive in nature or not. Genetic algorithm is applied on the rule set database for pruning best individuals. This process makes the decision faster as the search space of the resulting rule set is much compact when compared to the original data set. This makes IDS faster and intelligent. This method exhibits a high detection rate with low false positives. DARPA Dataset is used for initial training and testing purpose**.** The main drawback of the IDS is that it has to be trained for every new type of application and a lot of legitimate traffic may be classified as an attack.

Tsai and chou [18], Author said that data pre-processing is one of the most important steps in KDD or data mining. In data pre-processing dataset is made as much as clean so that it may use efficiently in learning process. However, since data pre-processing including feature selection or dimensionality reduction and data reduction is a very important stage for successful data mining, very few consider performing both tasks to examine the impact of data pre-processing on prediction performance. Bankruptcy prediction can be accomplished by data mining techniques. In this research genetic algorithm is being used for data pre-processing tasks, for data reduction and feature selection over a bankruptcy prediction dataset. Particularly, in this experiment different priorities of performing feature selection and data reduction are conducted. The result of the system shows that the support vector machine (SVM) classifier can attain the highest rate of accuracy if reduced data is provided.

Srinivasa et al [19], In this paper, author has proposed an intelligent query answering system based on rough sets and genetic algorithms. It is critical in database management system to optimize the queries and the complexity involved in finding optimal solutions has led to the development of heuristic approaches. To answer data mining query it has to need a random search over large databases. Due to the enormous size of the data set involved, simplification of model is necessary for giving quick answers of data mining queries. Classification and summarization of the datasets is done by rough set. Whereas, genetic algorithms are being used for answering association related queries and feedback for adaptive classification. Scalability of the system can be increase by building the summary table of rough set. Experimental results of the system justify with acceptable level of accuracy and speed.

Liu and Deng [20], In this research, author introduced an evolutionary approach to diagnose breast cancer by using artificial neural networks and genetic algorithm. Breast cancer is a most common disease and a frequent cause of death in women in the 35-55 year age group worldwide, or can say it is second type of the most common cancer in women worldwide. Here in this research, a hybrid approach is used to diagnosis of breast cancer, in this approach adoptive genetic algorithm is being used with artificial neural network. Because adoptive genetic algorithms have capability of strong macro search and global optimization, so it is used to optimize initial weights of the network. Adaptive genetic algorithm (AGA) was applied to evolve back propagation (BP) neural network, to achieve 98.9% classification accuracy and minimum standard deviation. The main disadvantage of BP is that it can easily trapped in a local minimum, to avoid this problem, genetic algorithm is used because it is good at global search, and therefore, in this research adaptive genetic algorithm is adopted to optimize the initial weights and thresholds of the BP.

Li et al [9], Said that after mining of partial ore body, movement and deformation of rock is a very complex physical and mechanical process. Time series prediction provides a method to master the dynamic law of strata and ground movement, and also provides an effective way to predict the movement dynamically. Here in this paper BP neural network was used for time series prediction. BP neural network has several limitation such as easily falling into local minimum, slowly convergence speed and

difficult to determine the initial weights etc. Genetic algorithm was used to optimize the initial neural network weight to avoid all problems coming in BP neural network. Then BP neural network was trained to establish a time series prediction model with samples of initial weights. To avoid the network falling into local minimum the initial network weight can be selected effectively to use BP neural network for mining subsidence time series prediction, thus the network forecasting performance can be improved effectively. This research provides a new method for dynamic mining subsidence prediction.

Pan et al [21], In this research, author introduces a new method of gas emission forecasting based on the optimized Radial Basis Function neural network. Here in this method, genetic algorithm is used to optimize the position of data centers, weights and widths of the RBF network, thus forming model is a GA-RBF model. The simulation results show that the improved RBF neural networks give the reliable and more accurate results with fast network training speed and good convergence rate. The method is found more efficient and feasible when compare with traditional RBF and BP networks.

Amin et al [22], In this research, author introduces a new method for heart disease prediction based on the neural network and genetic algorithm. All the existing systems predict heart diseases by using clinical dataset collected from complex tests conducted in pathology labs. None of the system predicts heart diseases based on risk factors such as diabetes, age, family history, hypertension, high cholesterol, alcohol intake, tobacco smoking, obesity or physical inactivity, etc. But the introduced system would give patients a warning about the probable presence of heart disease even before he goes for costly medical checkups in pathology labs. In this technique uses two most successful data mining tools, genetic algorithms and neural networks. In this hybrid system genetic algorithm is used for optimization of neural networks weights, so that the system may not fall in to local minimum. The learning process of derived system is fast, more stable and accurate as compared to back propagation. The system predicts the risk of heart disease with an accuracy of 89%.

III. CONCLUSION

From the above study we conclude that neural networks and genetic algorithms are two promising data mining tools widely used for classification and prediction in a complex data set. One of the main reason for using neural networks and genetic algorithms in data mining is robustness of solution and scalability of system. GA is global search optimization algorithm which leads to best and optimize solution for a complex problem, GA works on potential solution population which allows the elimination of weak individuals and favour the survival of the best ones. Neural networks are very good in classification task but there are two major disadvantages with neural networks. First is that the initialization of the neural networks weights is a blind process so it is not possible to initialize the globally optimized initial weights so there is risk that network output can run towards local optima which can affect the global solution. . The second problem with neural networks is that they are very slow in convergence and so it is possible that network may never converges. These problems of neural networks can be solved by using optimized initial weights, optimized by GA. Finally we conclude that hybrid Neural-Genetic approach is more efficient in data mining problems.

REFERENCES

[1] Jaiwei Han & Micheline Kamber "Data mining concept & techniques." *Morgan Kaufmann Publishers (Elsevier) 2001, ISBN* 1-55860-489-8.
[2] Frawley and Piatetsky-Shapiro, *Knowledge Discovery in Databases*: An Overview. (The AAAI/MIT Press, Menlo Park, C.A 1996.).
[3] Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P. (1996). From *Data Mining to Knowledge Discovery in Databases*.(AI Magazine, 17(3), 37-54.)
[4] Anil Jain k., Jianchang Mao and K.M. Mohiuddi, "Artificial Neural Networks: A Tutorial", *IEEE Computers*, (1996) pp.31-44.
[5] George Cybenk, "Neural Networks in Computational Science and Engineering", *IEEE Computational Science and Engineering*, (1996), pp.36-42.
[6] Rojas, "Neural Networks: a systematic introduction", *Springer-Verlag* (1996).
[7] R.P.Lippmann, "Pattern classification using neural networks," *IEEE Commun. Mag*, (1989), pp. 47–64.
[8] Simon Haykin, "*Neural Networks – A Comprehensive Foundation*", (Pearson Education 2001).
[9] Peixian LI, Zhixiang TAN, Lili YAN, Kazhong DENG, " Time series prediction of mining subsidence based on genetic algorithm neural network ", *International Symposium on Computer Science and Society*2011.
[10] Dr. Yashpal Singh, Alok Singh Chauhan, "Neural Networks in Data Mining", *Journal*

of Theoretical and Applied Information Technology 2005-2009.

[11] Dr. 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.

[12] Rudy Setiono, and Huan Liu"Effective Data Mining Using Neural Networks", *IEEE Transactions on Knowledge and Data Engineering, Vol. 8, No. 6*, December 1996.

[13] Taiwo Ayodele, Shikun Zhou, Rinat Khusainov"Email Classification Using Back Propagation Technique", *International Journal of Intelligent Computing Research (IJICR), Volume 1, Issue 1/2*, March/June 2010.

[14] Xianjun Ni "Research of Data Mining Based on Neural Networks", *World Academy of Science, Engineering and Technology*, 2008.

[15] Li Gao, Shangping Dai, Shijue Zheng, Guanxiang Yan"Using Genetic Algorithm for Data Mining Optimization in an Image Database", *Fourth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)* 2007.

[16] Wenxiang Dou, Jinglu Hu, Kotaro Hirasawa and Gengfeng Wu "Quick Response Data Mining Model Using Genetic Algorithm ", *SICE Annual Conference, The University Electro-Communications, Japan* 2008.

[17] K G Srinivasa, SaumyaChandra, Siddharth Kajaria, Shilpita Mukherjee"IGIDS: Intelligent Intrusion Detection System Using Genetic Algorithms", *IEEE* 2011.

[18] Chih-Fong Tsai and Jui-Sheng Chou"Data Pre-Processing by Genetic Algorithms for Bankruptcy Prediction", *IEEE* 2011.

[19] K G Srinivasa and Jagadish M, K R Venugopal, L M Patnaik"Data Mining based Query Processing using Rough Sets and Genetic Algorithms ", *Proceedings of the 2007 IEEE Symposium on Computational Intelligence and Data Mining* 2007.

[20] Lijuan Liu and Mingrong Deng "An Evolutionary Artificial Neural Network Approach for Breast Cancer Diagnosis", *Third International Conference on Knowledge Discovery and Data Mining* 2010.

[21] Yumin Pan and Weining Xue, Quanzhu Zhang,Liyong Zhao "A Forecasting Model of RBF Neural Network Based on Genetic Algorithms Optimization ", *Seventh International Conference on Natural Computation* 2011.

[22] Syed Umar Amin, Kavita Agarwal, Dr. Rizwan Beg, " Genetic Neural Network Based Data Mining in Prediction of Heart Disease Using Risk Factors ", *Proceedings of IEEE Conference on Information and Communication Technologies* (ICT 2013).