UDC 575
DOI: 10.2298/GENSR14031013V
Original scientific paper

A MICROARRAY GENE EXPRESSION DATA CLASSIFICATION USING HYBRID BACK PROPAGATION NEURAL NETWORK

M.VIMALADEVI[1*] and B.KALAAVATHI[2]

[1]Department of Computer Science and Engineering,
Hindustan Institute of Technology, Coimbatore, Tamilnadu, India
[2]Department of Computer Science and Engineering,
K.S.R. Institute for Engineering and Technology, Tiruchengode, Tamilnadu, India

Vimaladev M, and.B. Kalaavathi (2014): *A microarray gene expression data classification using hybrid back propagation neural network*- Genetika, Vol 46, No. 3, 1013-1026.

Classification of cancer establishes appropriate treatment and helps to decide the diagnosis. Cancer expands progressively from an alteration in a cell's genetic structure. This change (mutation) results in cells with uncontrolled growth patterns. In cancer classification, the approach, Back propagation is sufficient and also it is a universal technique of training artificial neural networks. It is also called supervised learning method. It needs many dataset for input and output for making up the training set. The back propagation method may execute the function of collaborate multiple parties. In existing method, collaborative learning is limited and it considers only two parties. The proposed collaborative function can perform well and problems can be solved by utilizing the power of cloud computing. This technical note applies hybrid models of Back Propagation Neural networks (BPN) and fast Genetic Algorithms (GA) to estimate the feature selection in gene expression data. The proposed research work examines many feature selection algorithms which are "fragile"; that is, the superiority of their results varies broadly over data sets. By this research, it is suggested that this is due to higher-order interactions between features causing restricted minima in search space in which the algorithm becomes attentive. GAs may escape from such minima by chance, because works are highly stochastic. A neural net classifier with a genetic algorithm, using the GA to select features for classification by the neural net and incorporating the net as part of the objective function of the GA.

Corresponding author: M.Vimaladevi, Assistant Professor, Department of Computer Science and Engineering, Hindustan Institute of Technology, Coimbatore Tamilnadu, India. E-Mail: vimaladevmm@gmail.com

Key words: Gene Expression Data, Back Propagation Neural Network, Genetic Algorithm, Feature Selection.

INTRODUCTION

Cancer is the state where the body cells lose their functions, start multiplying and dividing uncontrollably. A cancer cells accumulate and form tumors (bulks). With the mechanization in hospitals, an enormous quantity of data is collected. Though, human decision-making is frequently optimal, it is pitiable when there are vast amounts of data to be confidential. Medical data mining has great possible for exploring hidden patterns in the data sets of medical field. These patterns can be used for scientific diagnosis of cancer. Neural Networks are one of a lot of data mining analytical tools that can be exploiting to make predictions for medical data. BPN uses the gradient based approach which either teaches slowly or obtain strut with local minimum. Instead of using gradient-based learning techniques, one may apply the commonly used optimization methods such as Genetic Algorithms (GAs), Particle Swarm Optimization (PSO), Ant Colony optimization to find the network weights. Fast GA is a stochastic general search method, capable of effectively exploring large search spaces, and has been used with Back Propagation Network(BPN) for determining the a variety of parameters such as quantity of hidden nodes and hidden layers, select relevant feature subsets, the learning rate, the momentum, and initialize and optimize the network connection weights. This research presents the application of hybrid model that integrates Fast Genetic Algorithm and BPN for diagnosis of cancers.

LITERATURE SURVEY

The Evolutionary Algorithms (EAs) are global, parallel, search and optimization methods, established on the principles of natural selection and population genetics (ÜNAL *et al*., 2013). In all-purpose, any iterative, population based approach that uses selection and random variation to generate new solutions can be regarded as an EA. The evolutionary algorithms field has its origins in four landmark evolutionary approaches: Evolutionary Programming (EP), Evolution Strategies (ES), Genetic Algorithms (GA), and Genetic Programming (GP). The genetic algorithm was popularized and, as a result, the majority of control applications in the literature accept this approach.

Feed forward back-propagation neural network is an important dynamic network approach to predict the gene regulatory network from the time course gene expression data. (BARMA *et al*., 2013) found metabolic pathway of a real-life microarray time series gene expression data set of yeast cell cycle. First they clustered the gene set using a clustering method and calculate the center of each cluster. In this article, they present the metabolic pathway by constructing neural networks over the cluster-center matrix of microarray data with feed-forward back-propagation neural network approach. From the output of gene regulatory network they find metabolic pathway which leads to drug discovery in future.

The Spiking Wavelet Radial Basis Neural Network can be effectively used for the classification of gene expression data are discussed (CHANDRA *et al*., 2014). A new spiking function has been proposed in the non-linear integrate and fire model and it's inter spike interval is derived and used in the Wavelet Radial Basis Neural Network for the classification of gene expression data. The proposed model is termed as Spiking Wavelet Radial Basis Neural Network (SWRNN).

DNA methylation data produced from the Illumina's Infinium Human Methylation 450K Bead Chip platform, in an effort to correlate interesting methylation patterns with cancer

predisposition and in particular breast cancer and B-cell lymphoma. Specifically, feature selection and classification are exploited in order to select the most reliable predictive cancer biomarkers, and assess their classification power for discriminating healthy versus cancer related classes (VALAVANIS *et al*., 2013). The selected features, which could represented predictive biomarkers for the two cancer types, attained high classification accuracies when imported to a series of classifiers. The results support the expediency of the methodology regarding its application in epidemiological studies.

A hybrid approach in their articled which combines the advantages of fuzzy sets, Pulse Coupled Neural Networks (PCNNs), and support vector machine, in conjunction with wavelet-based feature extraction. An application of MRI breast cancer imaging has been chosen (HASSANIEN *et al*., 2012) and hybridization approaches have been applied to see their ability and accuracy to classify the breast cancer images into two outcomes: normal or non-normal.

A Rough Set (RS) based supporting vector machine classifier (RS_SVM) is proposed for breast cancer diagnosis (CHEN *et al*., 2011). In the proposed method (RS_SVM), RS reduction algorithm is employed as a feature selection tool to remove the redundant features and further improve the diagnostic accuracy by SVM. The effectiveness of the RS_SVM is examined on Wisconsin Breast Cancer Dataset (WBCD) using classification accuracy, sensitivity, specificity, confusion matrix and Receiver Operating Characteristic (ROC) curves.

The dimension reduction of DNA features in which relevant features are extracted among thousands of irrelevant ones through dimensionality reduction is addressed (BAI *et al*., 2014). This enhances the speed and accuracy of the classifiers. Principal Component Analysis (PCA) is a very powerful statistical technique to represent the d-dimensional data in a lower-dimensional space without any significant loss of information.

The effects of the parameters of parallel GAs on the quality of their search and on their efficiency are not well understood. This insufficient knowledge limits our ability to design fast and accurate parallel GAs that reach the desired solutions in the shortest time possible. The goal of this dissertation is to advance the understanding of parallel GAs and to provide rational guidelines for their design. The research reported here considered three major types of parallel GAs: simple master-slave algorithms with one population, more sophisticated algorithms with multiple populations, and a hierarchical combination of the first two types. The investigation of (CANTU-PAZ, 1999) formulated simple models that predict accurately the quality of the solutions with different parameter settings. The quality predictors were transformed into population-sizing equations, which in turn were used to estimate the execution time of the algorithms.

They established the Gene Expression Messy Genetic Algorithm (GEMGA)-a new generation of messy GAs that may find many applications in financial engineering by KARGUPTA and BUESCHER (1996). Contrasting other existing blackbox optimization algorithms, GEMGA straightly searches for relations among the associates of the search space.

During the learning process several passes are made over the training data set by the Genetic Algorithm and this makes it extensively I/O intensive and inappropriate. One technique to work out this problem is to build the model incrementally. VIVEKANANDAN and NEDUNCHEZHIAN (2010) suggested an incremental Genetic Algorithm that builds the rule based classification model in a fine granular manner by independently evolving tiny components based on the evolution of the data set which reduces the learning cost and thus building it scalable to large data sets.

They made an attempt to develop English character recognition system. (KAUR *et al*., 2011), describes the process of character recognition using the hybrid algorithm of Back

Propagation and Genetic Algorithm for the recognition of uppercase alphabets. In the survey, it is found that back propagation is although an efficient technique for training multilayer feed forward network.

A novel method based on a multi–objective genetic algorithm is investigated that evolves a near–optimal trade–off between Artificial Neural Network (ANN) classifier accuracy (sensitivity and specificity) and size (number of genes). This hybrid method (KEEDWELL *et al*., 2013), is shown to work on four well–established gene expression data sets taken from the literature.

They proposed a multi-objective Particle Swarm Optimization (PSO)-based algorithm that optimizes average node-weight and average edge-weight of the candidate subgraph simultaneously. The proposed algorithm is applied by (MANDAL *et al*., 2014) for identifying relevant and non-redundant disease-related genes from microarray gene expression data.

The recent effort in trying to understand the role of heterogeneity in cancer progression by using neural networks to characterize different aspects of the mapping from a cancer cells genotype and environment to its phenotype. (GERLEE *et al*., 2014) central premise is that cancer is an evolving system subject to mutation and selection, and the primary conduit for these processes to occur is the cancer cell whose behavior is regulated on multiple biological scales. The selection pressure is mainly driven by the microenvironment that the tumor is growing in and this acts directly upon the cell phenotype.

They utilized a fuzzy *c*-means clustering hybrid approach that combines support vector regression and a genetic algorithm. In this method, the fuzzy clustering parameters, cluster size and weighting factor are optimized and missing values are estimated. The proposed novel by (AYDILEK *et al*., 2013) hybrid method yields sufficient and sensible imputation performance results.

A hybrid M5′-Genetic Programming (M5′-GP) approach is proposed by (GARG *et al*., 2013) for empirical modeling of the FDM process with an attempt to resolve this issue of ensuring trustworthiness. This methodology is based on the error compensation achieved using a GP model in parallel with an M5′model.

The study was aimed at evaluating apoptotic potential of Br-oxph (*4-bromo-N,N-diethyl-5,5-dimethyl-2,5-dihydro-1,2-oxaphosphol-2-amine 2-oxide*) *in vitro*. (KOLEVA *et al*., 2014) provided a quantitative assessment of the apoptotic potential of Broxph in human lung carcinoma cells at concentrations corresponding to IC50 and 2xIC50 for 3 hours. Treatment with 2xIC50 significantly increased the amount of cytoplasmic DNA-fragments.

Fanconi anemia (FA) is a rare genetically heterogeneous disease characterized by developmental abnormalities, progressive bone marrow failure, and cancer susceptibility. (JOKSIĆ *et al*., 2013) examined spontaneous, diepoxybutane (DEB)- induced and radiation-induced sister chromatid exchanges (SCEs) in whole blood lymphocyte cultures of bone marrow failure (BMF) patients including Fanconi anemia, mothers of affected individuals, and healthy controls.

Assessment of genetic control, mode of inheritance, general and specific combining abilities and effect of drought stress on genetic parameters of harvest index and biological yield traits in bread wheat were achieved by (GOLPARVAR, 2014) using Diallel mating design.

MATERIALS AND METHODS

This section illustrates the proposed feature selection of Hybrid Fast Genetic Algorithm. The main aspects of the algorithm namely, Back Propagation Neural Network, Genetic Algorithm and Modified Successive Feature Selection are discussed.

FEATURE SELECTION

Feature subset selection is of enormous significance in the domain of data mining. The elevated dimension data creates testing and training of general classification methods complicated. Feature selection is significant pre-processing method to eliminate irrelevant and redundant data. It can be applied in both unsupervised and supervised learning. In supervised learning, feature selection aspires to maximize classification accuracy. The objective of feature selection for unsupervised learning is to discover the smallest feature subset that best uncovers clusters form data according to the preferred criterion. Successive Feature Selection (SFS) procedure, a set of $x \leq 10$ features is practiced one at a time that the value of x is in use due to recollection constrictions and it is experimentally established that the appropriate values of x is equal to or lower than 10. The production is the rank of features. In the successive level that the feature is dropped one at a time and a subset of features are gained. The classification accuracy is evaluated using classifiers, and the best subset of features is processed to the next level. There could be more than one top subset of features in a given level. From the figure, it is deduced that a feature is dropped in level 1 gives four different subsets of features. The best set in level 1, s$= \{x_1, x_2, x_4\}$ which is selected for level 2. In a comparable way a feature is dropped from the best set of features of level 1 into level 2, which gives three different subsets of features. The best subsets in level 2 are $\{x_2, x_4\}$ and $\{x_1, x_2\}$ and their classification accuracies are the same and higher than those of other subsets and the best subset in level 3 is$\{x_2\}$.This process is terminated when all the features are ranked. Two ranked sets are obtained in SFS: namely $R_1 = \{x_2, x_4, x_1, x_3\}$ and$R_2 = \{x_2, x_1, x_4, x_3\}$.

Modified Successive Feature Selection

In the SFS two ranked SFS are obtained, which indicate that $x_2$is the top-ranked feature and x3 is the bottom ranked or least important feature. Want to select the three top-ranked features, then the result will be $F_1 = \{x_2, x_4, x_1\}$ and $F_2 = \{x_2, x_1, x_4\}$. If the order of features is not important, then instead of using two sets, F1 and F2, select a common top three ranked features from the set

$$F_k = F_1 \cup F_2 = \{x_1, x_2, x_4\}$$

Now, the Gene ranking can be found by Mean and Standard Deviation, first the mean have to be taken for the first top three ranked features. Then the standard deviation has taken for common top three ranked features. Then the Gene ranking is calculated by maximum value of Mean to the Standard deviation.

The Modified Successive Feature Algorithms are given as

Step1: Find the set of features from $F_1 = \{x_2, x_4, x_1\}$and$F_2 = \{x_2, x_1, x_4\}$.

Step2: Top three genes are selected by applying intersection on$F_1$and F_2, that is,

$$F_k = F_1 \cup F_2 = \{x_1, x_2, x_4\}$$

Step3: The gene rankings are found out using three features

Step4: Gene Ranking=$\frac{F_k - mean(F_k)}{standarddeviationofF_k}$ K denotes the three features value (i.e., 1, 2, 3)

Step5: Find out the maximum value of gene ranking. The Modified successive feature selection algorithm investigates the importance gene and the procedure is given in figure 1.

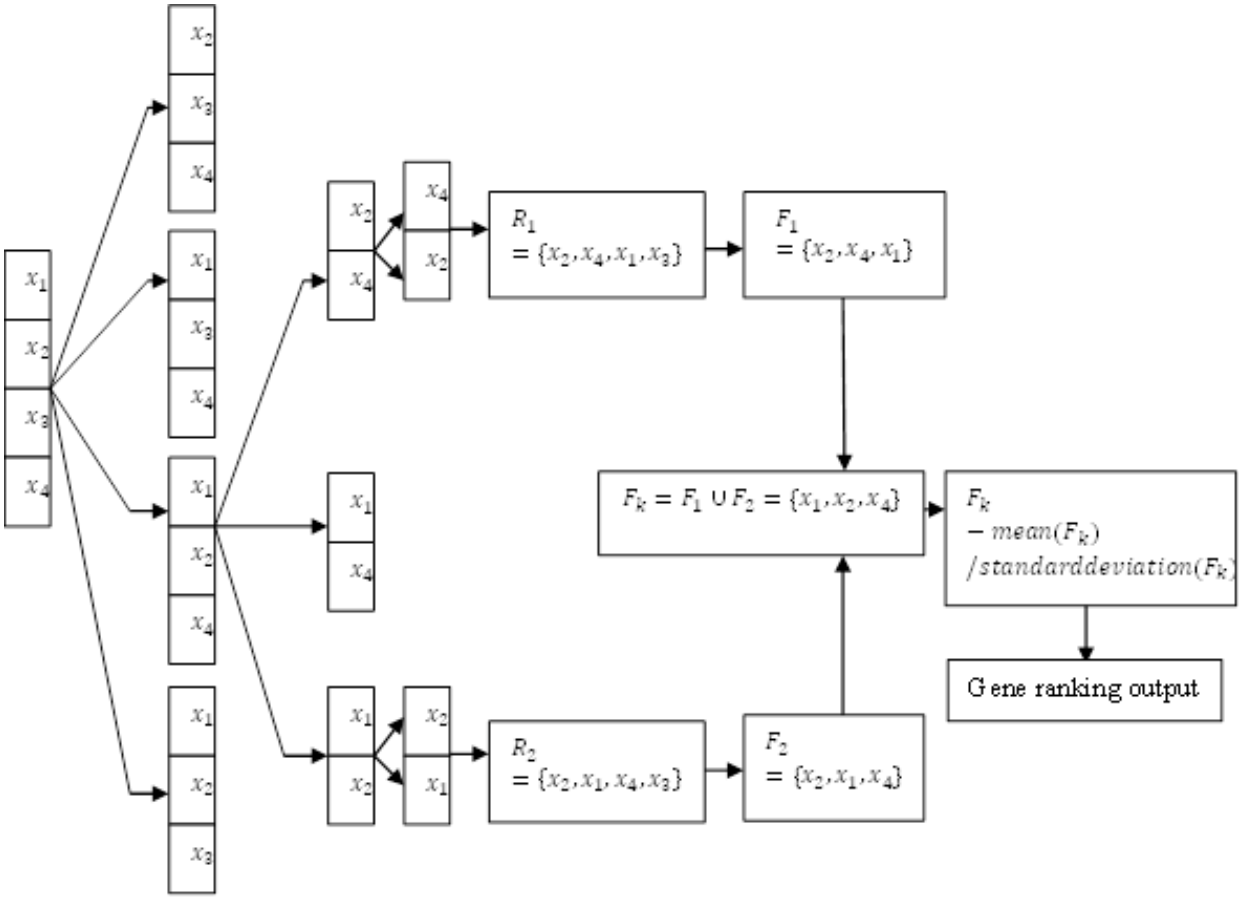

Figure 1 Modified Successive Feature Selection

BACKPROPAGATION NEURAL NETWORKS

BPN is an information-processing prototype that is inspired by the method biological nervous systems, such as the brain, process information. The key constituent of this paradigm is the novel structure of the information processing system. It is self-possessed of a large number of highly interconnected processing elements (neurons) working in agreement to find out the solution specific problems. Developing a neural network involves first training the network to carry out the needed computations. The feed-forward neural network architecture is commonly used for supervised learning. Feed-forward neural networks contain a set of layered nodes and weighted connections between nodes in adjacent layers. Feed-forward networks are frequently trained using a back propagation-learning scheme. Back propagation learning works by making modifications in weight values starting at the output layer then moving backward through the hidden layers of the

network. Neural networks have been criticized for their poor interpretability, since it is difficult for humans to interpret the representative meaning behind the learned weights.

$$E = {}^{1}/_{2} \sum (T_i - O_j)^2 \quad (1)$$

Advantages of neural networks, however, include their high tolerance to noisy data as their ability to classify patterns on which have not been trained.

GENETIC ALGORITHM

GA is an optimization techniques inspired by natural selection and natural genetics. Contrasting many search algorithms, which achieve a local, greedy search, GA is a stochastic general search method, competent of effectively exploring large search spaces. A genetic algorithm is mainly composed of three operators: reproduction, crossover, and mutation. As a first step of GA, an initial population of individuals is generated at random or heuristically. The individuals in the genetic space are called chromosome. The chromosome is a collection of genes where genes can generally be represented by different methods like binary encoding, value encoding, permutation encoding and tree encoding. Gene is the basic building block of the chromosome. Locus is the position of particular gene in the chromosome.

In every generation, the population is appraised using fitness function. Next approaches the selection process, where in the high fitness chromosomes are used to eliminate low fitness chromosomes. The commonly used methods for reproduction or selection are Roulette-wheel selection, Boltzmann selection, Tournament selection, Rank selection and Steady-state selection. But selection unaccompanied does not fabricate any new individuals into the population. Hence selection is followed by crossover and mutation operations. Crossover is the process by which two-selected chromosome with high fitness values exchange part of the genes to generate new pair of chromosomes. The crossover tends to facilitate the evolutionary process to progress toward potential regions of the solution space. Different types of crossover by and large used are one point crossover, two-point crossover, uniform crossover, multipoint crossover and average crossover. Mutation is the random change of the value of a gene, which is used to prevent premature convergence to local optima. Major ways that mutation is accomplished are random bit mutation, random gene mutation, creep mutation, and heuristic mutation. The new population generated undergoes the further selection, crossover and mutation till the termination criterion is not satisfied. Convergence of the genetic algorithm depends on the various principles like fitness value achieved or number of generations.

HYBRID MODEL OF FAST GA-BPN

Back propagation learning works by making modifications in weight values starting at the output layer then moving backward through the hidden layers of the network. BPN uses a gradient method for finding weights and is horizontal to direct to troubles such as local minimum problem, slow convergence pace and convergence unsteadiness in its training procedure. Contrasting numerous search algorithms, which execute a local, greedy search, GAs performs a global search. GA is an iterative procedure that consists of a constant-size population of individuals called chromosomes, every one represented by a finite string of symbols, known as the genome, encoding a possible solution in a given problem space. The GA can be employed to improve the

performance of BPN in different ways. GA is a stochastic general search method, capable of effectively exploring large search spaces, which has been used with BPN for determining the number of hidden nodes and hidden layers, select relevant feature subsets, the learning rate, the momentum, and initialize and optimize the network connection weights of BPN. GA has been used for optimally designing the ANN parameters including, ANN architecture, weights, input selection, activation functions, BPN types, training algorithm, numbers of iterations, and dataset partitioning proportion.

The Hybrid Fast GA-BPN has been used in the miscellaneous applications. GA has been used to search for optimal hidden-layer architectures, connectivity, and training parameters (learning rate and momentum parameters) for BPN for predicting community-acquired pneumonia among patients with respiratory complaints. GA has been used to initialize and optimize the connection weight of BPN to improve the presentation BPN and is applied in a medical difficulty for predicting stroke disease. GA has been used to optimize the BPN parameters namely: learning rate, momentum coefficient, Activation function, Number of hidden layers and number of nodes for worker assignment into Virtual Manufacturing Cells (VMC) application. GA-BPN model has been experimented for of study of the heat transport characteristics of ananofluid thermosyphon in a magnetic field where, GA is used to optimize the number of neurons in the hidden layer, the coefficient of the learning rate and the momentum of BPN.

The present work demonstrates the application of fast GA for initializing and optimizing the connection weights of BPN. The leading step of the GA is demonstration of the chromosome. For the BPN with single hidden layer with m nodes, n input nodes and p output nodes the number of weights to be computed is given by (n+p)* m. Each chromosome is finished up of (n+p)*m number of genes.

Genes are represented by real number encoding method. The original population is a set of Chromosomes, which is generated randomly. Fitness of each chromosome is computed by minimum optimization method. Fitness is given by condition (Ci) = 1/ E for each chromosome of the population, where E is the error computed as root mean square error at the output layer as shown in equation 1, where summation is performed overall output nodes pj and tj is the desired or target value of output oj for a given input vector.

$$E = \frac{1}{2}\sum_{p}\sum_{j}\left(t_{pj} - o_{pj}\right)^2 \tag{2}$$

Once fitness is computed for the all the chromosomes, the best-fit chromosomes replace the worst fit chromosomes. Further crossover step is experimented using single point crossover, two-point crossover and multi point crossover. In addition a new type of crossover called mixed crossover has been used where for the given number of generation M, first 60% generation work applied multipoint crossover, followed by next 20% generation using two point crossover and remaining using one point crossover. Finally mutation is applied as the last step to generate the new population. The new population is given as input to PN to compute the fitness of each chromosome, followed by process of selection, reproduction, and cross over and mutations to generate the next population. This process is repeated till more or less all the chromosomes converge to the same fitness value. The weights represented by the chromosome in the final converged population are the optimized connection weights of the BPN. The working of Hybrid Fast GA-BPN for optimizing connection weights is shown in figure 2.

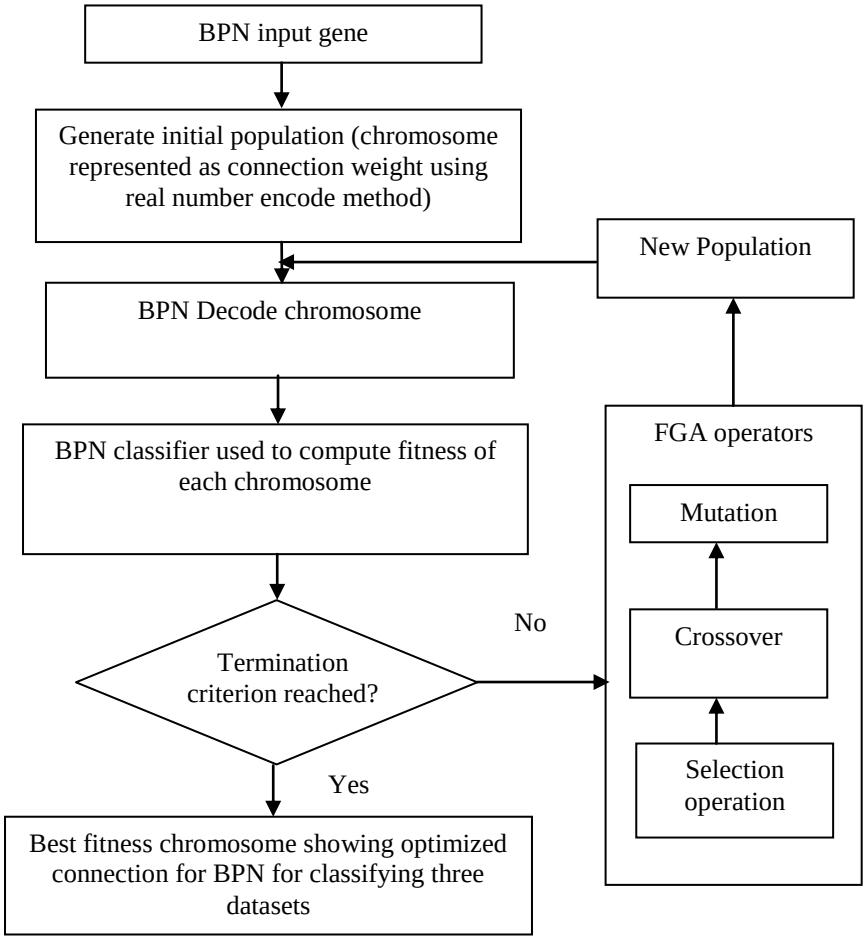

Figure 2. Hybrid Model of Fast GA-BPN

a) Hybrid Fast GA-BPN Algorithm:

The hybrid algorithm of Back Propagation and Fast Genetic Algorithm will be designed to train and test the network. The algorithm follows the steps:

1) **Coding**: First the initial values are to be in some type of coding. Real number coding system is adopted in this work. This network is of configuration l-m-n.Number of weights calculated as= ((l+n)*m)

2) **Weight Extraction**: To determine the fitness value for each chromosome show extract weights from each of chromosome.

3) **Fitness**: The fitness function must be devised for each problem to be solved. First errors are calculated as:

$$w_k = +x_{kd+2}10^{d-2} + x_{kd+3}10^{d-3} + \cdots\cdots + x_{(k+1)d}/10^{d-2} \quad if\ 5 \leq x_{kd+1} \leq 9 \tag{3}$$

$$w_k = +x_{kd+2}10^{d-2} + x_{kd+3}10^{d-3} + \cdots\cdots + x_{(k+1)d}/10^{d-2} \quad if\ 0 \leq x_{kd+1} \leq 5 \tag{4}$$

4) **Reproduction**: In this step formation of mating pool is done. Mating pool is formed by excluding that chromosome with least fitness value and replacing it with a duplicate copy of the chromosome reporting the highest fitness value. The three different operators of genetic algorithm are applied to update the population in the reproduction.

- Selection
- Crossover
- Mutation

$$E_1 = (T_{ij} - O_{ij})^2 + (T_{ij} - O_{ij})^2 \tag{5}$$

$$E = \sqrt{E_1 + E_2 + \cdots\cdots + E_n}/n \tag{6}$$

$$F = \frac{1}{E} \tag{7}$$

5) **Convergence**: A population is said to be converged when 95% of individuals constituting the population share same fitness value.

The final outputs given by the algorithm are the final weights to be adjusted for the neural network. So in this hybrid approach the calculate robustness function using Fast Genetic approach instead of mass as in Back propagation algorithm.

RESULTS

In this research, demonstrates about the experimental setup of data sets.

Datasets

Three DNA microarray gene expression data sets namely, SRBCT, Leukemia and Lymphoma are used for experimentation purposes. Their performance in terms of classification accuracy using only the features is very promising.

A) SRBCT Dataset: The small round blue-cell tumor dataset consists of 50 samples, each containing 2,308 genes. This is a classification problem. The tumors are Burkitt's lymphoma (BL), the Ewing Family of tumors (EWS), NeuroBlastoma (NB), and RhabdoMyoSarcoma (RMS). Thereare 63 samples for training and 20 samples for testing. The training set consists of 8, 23, 12, and 20 samples of BL, EWS,NB, and RMS, respectively. The test set consists of 3, 6, 6, and5 samples of BL, EWS, NB, and RMS, respectively.

B) Leukemia Dataset: It consists of 72 samples: 25 samples of Acute Myeloid Leukemia (AML) and 47 samples of Acute Lymphoblastic Leukemia (ALL). The source of the gene expression measurements is taken from 63 bone marrow samples and 9 peripheral blood samples. Gene expression levels in these 72 samples are measured using high density oligonucleotide microarrays. Each sample contains 7129 gene expression levels.

C) Lymphoma Dataset: B cell Diffuse Large Cell Lymphoma (B-DLCL) is a heterogeneous group of tumors, based on significant variations in morphology, clinical

presentation, and responseto treatment. Gene expression profiling has revealed two distinct tumor subtypes of B-DLCL: germinal center B cell-like DLCL and activated B cell-like DLCL. Lymphoma dataset consists of 24 samples of germinal center B-like and 23samples of activated B-like.

Table 1 Summary of the Data Sets Used in the Experimentation

Dataset	Training set	Test set
SRBCT	63	20
Leukemia	72	52
DLBCL	77	21

Testing Accuracy and Execution Time

In table2 and 3 shows the accuracy and execution time for both BPN and Hybrid Fast GA-BPN are given in tabulation.

Table 2 Comparison of BPN and Hybrid Fast GA-BPN for Testing Accuracy

DATASET	NO OF GENE COMBINATION	ACCURACY (%)	
		BPN	Hybrid GA-BPN
LYMPHOMA	100, 2	59.42	85.65
LEUKEMIA	100, 3	65.23	89.33
SRBCT	100,4	72.31	91.27

The table 2 represents the accuracy and execution time for gene expression data using the Fast GA technique. The comparisons of BPN and hybrid approaches are evaluated using three datasets lymphoma, leukemia and SRBCT. The hybrid BPN is more proficient performance than the existing technique. The SRBCT are more sufficient and effective than other two datasets. Using SRBCT dataset the numbers of gene data are higher than other two datasets but it shows the accuracy and execution time is better in this performance.

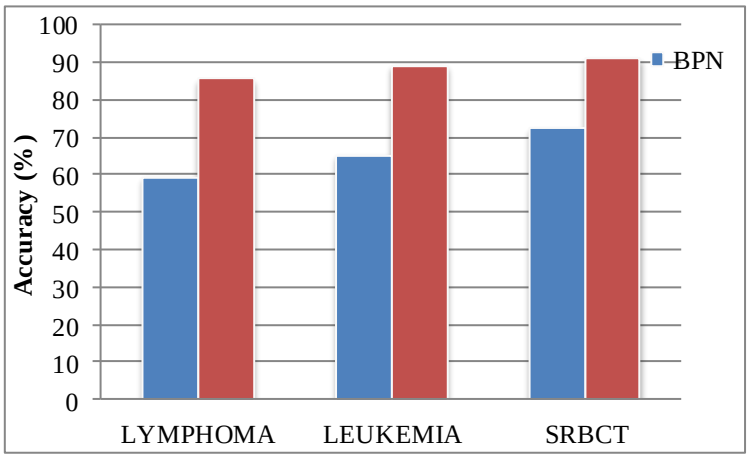

Figure 3 Shows the accuracy for BPN and Hybrid Fast GA-BPN

The accuracy value represents the better outcome when compared to the BPN and Hybrid Fast GA-BPN techniques. Fig3 shows the accuracy value for BPN and Hybrid Fast GA-BPN. Compared to BPN the accuracy value is higher in hybrid approach.

Table 3 Execution time for the comparison of BPN and Hybrid Fast GA-BPN

DATASET	NO OF GENES COMBINATION	EXECUTION TIME (SECONDS)	
		BPN	Hybrid GA-BPN
LYMPHOMA	100, 2	53	30
LEUKEMIA	100, 3	49	23
SRBCT	100,4	37	20

Table 3 shows the execution time for both BPN and Hybrid Fast GA-BPN. The running process are utilize very low in proposed hybrid approaches. The three datasets are exploring the outcome of execution time for gene expression data.

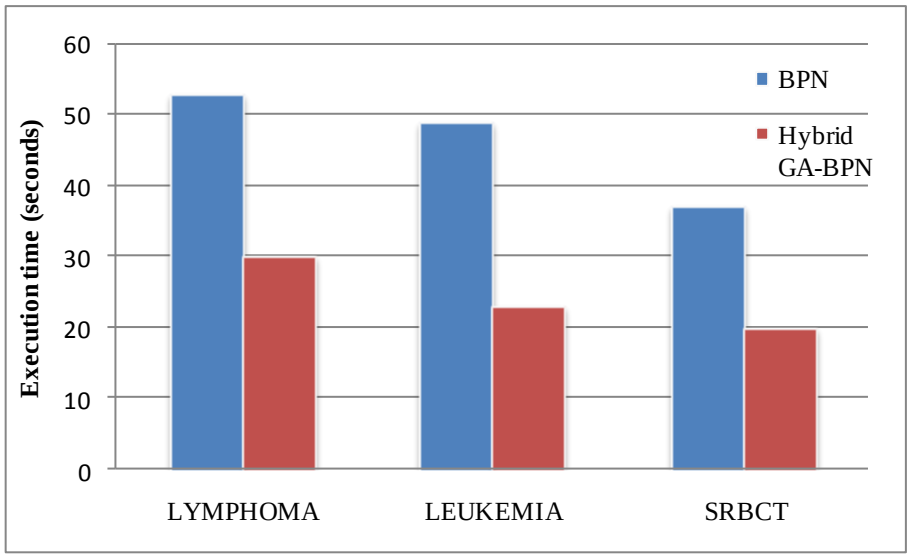

Figure 4 shows the execution time of BPN and Hybrid Fast GA-BPN

In Fig4 represents the execution time for comparable BPN and Hybrid Fast GA-BPN for gene expression data. In Hybrid Fast GA-BPN technique execution time is smaller than the BPN.

CONCLUSION

The high dimensionality of microarray gene expression data creates the need for methods which automatically detect interesting structures in the data. In this research, introduced mathematical criterion that characterizes the cancer subtypes represented in several gene expression data sets and have demonstrated an algorithm that, by employing this criterion, recovers these subtypes without using prior knowledge. In many feature selection algorithms (e.g., individual ranking and forward selection schemes), the gene selection is biased toward the highest

ranking feature. However, low-rank genes, if appropriately selected in a subset, can exhibit better classification performance. The proposed Hybrid Fast GA-BPN algorithm explores this phenomenon and provides a method to investigate important genes. It is observed that the algorithm finds a small gene subset that provides high classification accuracy on several DNA microarray gene expression data sets. In this research represented methodology will encourage future work should test the approach used for other similar tasks or other related data sets to estimate its capability to produce a similar accuracy with improved GA for detecting significant ratio in cancer classification.

Received July 25th, 2014
Accepted October 12th, 2014

REFERENCES

AYDILEK, I.B., and A. ARSLAN (2013): A hybrid method for imputation of missing values using optimized fuzzy< i> c</i>-means with support vector regression and a genetic algorithm. Information Sciences 233: 25-35.

BAI, A., and A. PRADHAN (2014): Feature Extraction and Classification of Microarray Cancer Data Using Intelligent Techniques. In Intelligent Computing, Networking, and Informatics. pp. 1277-1284. Springer India.

BARMAN, B., and A. MUKHOPADHYAY (2013): Computational Modeling of Metabolic Pathway through Construction of GRN within Clusters of Yeast Cell Cycle data.

CHANDRA, B., and K. V. NARESH BABU (2014): Classification of Gene Expression Data Using Spiking Wavelet Radial Basis Neural Network. Expert systems with applications 41 (4): 1326-1330.

CHEN, H.L., B. YANG, J. LIU, and D.Y. LIU (2011): A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis. Expert Systems with Applications 38(7): 9014-9022.

GARG, T.L., and SAVALANI (2013): A hybrid\ text {M} 5^\\ prime-genetic programming approach for ensuring greater trustworthiness of prediction ability in modelling of FDM process. Journal of Intelligent Manufacturing): 1-17.

GERLEE, P., E. KIM, and A.R.A ANDERSON (2014): Bridging scales in cancer progression mapping genotype to phenotype using neural networks. In Seminars in cancer biology. Academic Press.

GOLPARVAR, A. R. (2014): Genetic control and heredity of harvest index and biological yield in bread wheat (Triticum aestivum L.). Genetika, *46(1):* 43-48.

HASSANIEN, A.E., and T.H. KIM (2012): MRI Breast cancer diagnosis approach using support vector machine and pulse coupled neural networks. Journal of Applied Logic-Elsevier.
http://www.biomedcentral.com/content/supplementary/1471-2105-13-178-S1.docx

JOKSIĆ, I., S. PETROVIĆ, A. LESKOVAC, J. FILIPOVIĆ, M. GUČ-ŠĆEKIĆ, D. VUJIĆ, and G. JOKSIĆ (2013): Enhanced frequency of sister chromatid exchanges induced by diepoxybutane is specific characteristic of Fanconi anemia cellular phenotype. Genetika *45 (2):* 393-403.

KARGUPTA, H., and K. BUESCHER (1996): The gene expression messy genetic algorithm for financial applications. In Computational Intelligence for Financial Engineering. Proceedings of the IEEE/IAFE 1996 Conference on. pp. 155-161. IEEE.

KAUR, R., and B. SINGH (2011): A hybrid neural approach for character recognition system. Int J Comput Sci Inf Technol 2 (2): 721-726.

KEEDWELL, E.D., and A. NARAYANAN (2013): Gene expression rule discovery and multi–objective ROC analysis using a neural–genetic hybrid. International journal of data mining and bioinformatics 7(4): 376-396.

KOLEVA, V., A. DRAGOEVA, M. DRAGANOV, L. MELENDEZALAFORT, A. ROSATO, N. UZUNOV, and D. ENCHEV (2014): Inhibition of growth and induction of apoptosis in human lung cancer cells by Br-oxph. Genetika46 (1): 1-10.

MANDAL, M., and A. MUKHOPADHYAY (2014): A Graph-Theoretic Approach for Identifying Non-Redundant and Relevant Gene Markers from Microarray Data Using Multiobjective Binary PSO. PloS one 9 (3): e90949.

ÜNAL, M., A. AK, V. TOPUZ and H. ERDAL (2013): Genetic Algorithm. In Optimization of PID Controllers Using Ant Colony and Genetic Algorithms. Springer Berlin Heidelberg. pp. 19-29.

VALAVANIS, I.NIS, E. SIFAKIS, P. GEORGIADIS, S. KYRTOPOULOS, and A. CHATZIIOANNOU (2013): Derivation of Cancer Related Biomarkers from DNA Methylation Data from an Epidemiological Cohort. In Engineering Applications of Neural Networks. Springer Berlin Heidelberg. pp. 249-256.

VIVEKANANDAN and NEDUNCHEZHIAN (2010): A Fast Genetic Algorithm for Mining Classification Rules in Large Datasets. International Journal on Soft Computing (IJSC) *1(1):* 10-20.

KLASIFIKACIJA PODATAKA EKSPRESIJE GENA DOBIJENIH METODOM *MICROARRAY* KORIŠĆENJEM HIBRIDNOG *BPN* METODA *(BACK PROPAGATION NEURAL* NETWORK)

M.VIMALADEVI[1*], B.KALAAVATHI[2]

[1]Odeljenje za nauku o kompjuterima i inženjering, Hindustan Institut za tehnologiju, Coimbatore, Tamilnadu, Indija

[2] Odelenje za nauku o kompjuterima i inženjering, K.S.R. Institut za inženjering i tehnologiju Tiruchengode, Tamilnadu, Indija

Izvod

U klasifikaciji kancera metod *BPN* (*Back Propagation neural network)* je dovoljna I univerzalna tehnika za aktivnost veštačkih neuralnih mreža. Ova tehnika ima I naziv nadgledani metod učenja. Može da vrši funkciju kolaborativnih multiplih delova. Postojeća tehnika je ograničena jer razmatra samo dva dela. Predložena kolaborativna tehnika omogućava rešenje problema. Ona koristi hibridne modele *BPN i* brzi Genetički Algoritam (*GA*) u cilju utvrđivanja buduće selekcije podataka o ekspresiji gena. Ova istraživanja obuhvataju mnogo osobina selekcije algoritama, koji su fragilni (oseljivi) i superiornost rezultata njihovim korišćenjem široko varira u okviru grupa podataka. Dobijeni rezultati ukazuju da se to događa zbog interakcija višeg reda između osobina uzrokujući ograničeni minimum u sferi ispitivanja u k algoritam postaje atentativan.

Primljeno 25. VII 2014.

Odobreno 12. X. 2014.