ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 11, November 2014

Cancer Classification using Hybrid Fast Particle Swarm Optimization with Backpropagation Neural Network

M.Vimaladevi[1] , Dr.B.Kalaavathi[2]

Assistant Professor,Dept of Computer science and Engineering Hindusthan Institute of Technology,

Coimbatore, India[1]

Professor & Head, Dept of Computer science and Engineering, K.S.R.Institute for Engineering and Technology,

Tiruchengode, India[2]

Abstract : Cancer any type of malignant growth or tumor, caused by abnormal and uncontrolled cell division: it may spread through the lymphatic system or blood stream to other parts of the body .Cancer classification is known to have the keys for addressing the problems based on cancer diagnosis and drug discovery. The proposed method of DNA microarray technique has made continuous processing thousands of gene expressions possible. Using gene expression data, the researchers have started the performance to explore the possibilities of cancer classification. The various methods have been introduced in recent years with promising results. But there are still a lot of problem which need to be addressed and understood. The performance of combining the genetic algorithm and Hybrid Fast PSO-BPN method is used to solve the optimization problems. Hybrid Fast PSO-BPN method is used to improve the accuracy and better convergence rate of genetic algorithm and this method is better for local search. This proposed method is used to overcome from the problem of computational difficulties occur by ill-condition of the square penalty function. The experimental result shows that this proposed method is better in accurate result with less execution time.

Keywords: Microarray Gene Expression Data, Feature Selection, Gene Ranking, Genetic Algorithm, and Hybrid Fast PSO-BPN.

I INTRODUCTION

In medical field cancer is one of the major researches. Accurate prediction of various tumor types has huge value in offering better treatment and toxicity minimization on the patients. There are several limitations in previous cancer classification. It has been recommended that specifications of therapies according to tumor types differentiated by pathogenetic patterns may increase the efficacy of the patients.

DNA microarrays performs biologist to calculate the expression of thousands of genes continuously on a small chip. These microarrays perform the large amount of data and proposed methods are needed to analyze them. The hybrid method of fast particle swarm optimization with BPN method is used to classify gene expression data recorded on DNA microarrays. DNA microarrays can be used to calculate changes in expression levels of genes in various biological conditions. The proposed method may perform the better performance.

This new formulation is based on the reduction of the number of variables (number of generators) and elimination of the equality and inequality constraints, thus the transformation of the constrained non linear programming problem to an unconstrained one. The new unconstrained objective function is minimized by Hybrid Fast PSO-BPN method.

The review of this research is organized as follows. Section 2 summarizes the concepts and literature survey. Section 3 discusses the proposed method, and section 4 provides the experiments with high accuracy. Finally, Section 5 presents the conclusions of the work.

II LITERATURE SURVEY

A DNA microarray can track the expression levels of thousands of genes simultaneously. Previous research has demonstrated that this technology can be useful in the classification of cancers. Cancer microarray data normally contains a small number of samples which have a large number of gene expression levels as features are given by Wang *et al* (2005).

Duan *et al* (2005) proposee a new feature selection method that uses a backward elimination procedure similar to that implemented in support vector machine recursive feature elimination (SVM-RFE).

Statnikov *et al* (2005) proposed that a cancer diagnosis is one of the most important emerging clinical applications of gene expression microarray technology. Classification of patient samples is an important aspect of cancer diagnosis and treatment. The learning machine has been successfully applied to microarray cancer diagnosis problems. However, one weakness of the SVM is that given a tumor sample, it only predicts a cancer class label but does not provide any estimate of the underlying probability Zhu *et al* (2004).

Weigelt *et al* (2010) development of microarrays and the ability to perform massively parallel gene expression analysis of human tumours were received with great excitement by the scientific community. Simultaneous multiclass classification of tumor types is essential for future clinical implementations of microarray-based cancer diagnosis. In this study, we have combined genetic algorithms (GAs) and all paired Support Vector Machines (SVMs) for multiclass cancer identification are done by Peng *et al* (2003).

The development of microarray-based high-throughput gene profiling has led to the hope that this technology could provide an efficient and accurate means of diagnosing and classifying tumors, as well as predicting prognoses and effective treatments are given by Liu *et al* (2005). The problem of feature selection is a difficult combinatorial task in machine learning and of high practical relevance, e.g. in bioinformatics. Genetic Algorithms (GAs) offer natural ways to solve this problem are shown by Frohlich *et al* (2003).

The gene expression data obtained from microarrays have shown useful in cancer classification. DNA microarray data have extremely high dimensionality compared to the small number of available samples. Chen *et al* (2003) propose a novel system for selecting a set of genes for cancer classification. Shah *et al* (2007) propose gene expression data sets for ovarian, prostate, and lung cancer were analyzed in this research. This integrated algorithm involves a genetic algorithm and correlation-based heuristics for data preprocessing and data mining for making predictions.

Chen *et al* (2014) suggested to achieved efficient gene selection from thousands of candidate genes that can contribute in identifying cancers, this study aims at developing a novel method utilizing particle swarm optimization combined with a decision tree as the classifier. The study also compares the performance of their proposed method with other well-known benchmark classification methods (support vector machine, self-organizing map, and backpropagation neural network).

Haung *et al* (2012), compared the Particle Swarm Optimizer (PSO) based artificial neural network (ANN), the adaptive neuro-fuzzy inference system (ANFIS), and a case-based reasoning (CBR) classifier with a logistic regression model and decision tree model. Sahab *et al* (2005) presents a two-stage hybrid optimization algorithm based on a modified genetic algorithm. In the first stage, a global search is carried out over the design search space using a modified GA.

Sadoughi *et al* (2014), presents by combining the principles two algorithm, they proposed a new but simple hybrid algorithm called BPNN_PSO. Their novel algorithm optimizes BPNN with PSO and reduces computational time of the training phase of BPNN. Khan *et al* (2012) suggests that to show the superiority (time performance and quality of solution) of the new meta heuristic bat algorithm (BA) over other more "standard" algorithms in neural network training. In this work we tackle this problem with algorithms, and aims to over a set of results that could hopefully foster future comparisons by using a standard dataset.

III METHODOLOGY

This section shows the methodology for proposed feature selection of hybrid fast particle swarm optimization with BPN method.

3.1 Gene Ranking

Gene expression profiles should be properly preprocessed before analysis as pre-requisite, including background correction, normalization and summarization. Instead of the exact values, ranks of gene expression levels are used in the following procedure. A ranked list of genes was obtained first by sorting the microarray probe-set identifiers according to the different expression values (count or ratio). In gene ranking technique is selecting genes using feature-ranking filters

(1) Use a filter to rank all the genes in the data.

(2) Choose the first n-1 genes as the best feature subset.

3.2 Feature Selection

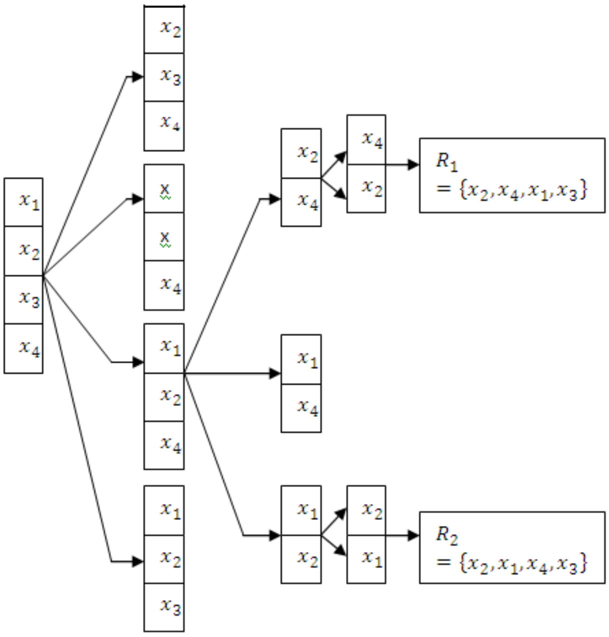

Fig. 1 Successive Feature Selection

Feature Selection for the classification of cancer data means discovering feature values and profiles of diseased and healthy samples. It also means using this knowledge to predict the state of new samples. Feature selection is the process of choosing a subset of input variables by eliminating irrelevant features. The elimination of irrelevant features reduces the dimensionality of data. It may also allow learning algorithms to operate faster and more effectively. Feature selection is an active research area in machine learning, pattern recognition, and data mining.

Feature selection consists of four phases: feature subset generation, feature subset evaluation, stopping criterion, and validation. There are different methods of feature selection based on search strategies and evaluation functions. The basic approach to subset generation is to start with an empty set and then add features to it based on the evaluation criteria.

This Successive Feature Selection (SFS) procedure is completed when all the features are ranked. Two ranked

]

sets are achieved in SFS: that is R1= {x2, x4, x1, x3} and R2= {x2, x1, x4, x3}

3.3 Fast Genetic Algorithm with BPN

Genetic Algorithms (GAs) are a search method which is based on principles of natural selection and genetics. GAs instructs the decision variables of a particular problem into length strings of alphabets of particular cardinality. The characters which are proper solutions to the problem are denoted to as chromosomes, the alphabets are denoted as genes and the values of genes are known alleles. For illustration, the problem like traveling salesman problem, a chromosome represents a route, and a gene may represent a city. GAs works with coding of parameters in the traditional optimization techniques. 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. Similar to the GA, the PSO algorithm is a global algorithm, which has a strong ability to find global optimistic result. Thus, combining PSO with backpropagation neural network in gene expression data are much effective than the GA-BPN.

3.4 Hybrid Fast Particle Swarm Optimization with Backpropagation Neural Network

By combining the PSO with the BPN, a new algorithm referred to as PSO–BPN hybrid fast algorithm is formulated in this research. The fundamental idea for this hybrid algorithm is that at the beginning stage of searching for the optimum, the PSO is employed to accelerate the training speed. When the fitness function value has not changed for some generations, or value changed is smaller than a predefined number, the searching process is switched to gradient descending searching according to this heuristic knowledge.

The Hybrid Fast PSO–BPN algorithm's searching process is also started from initializing a group of random particles. First, all the particles are updated, until a new generation set of particles are generated, and then those new particles are used to search the global best position in the solution space. Finally the BP algorithm is used to search around the global optimum. In this way, this hybrid algorithm may find an optimum more quickly. The procedure for this hybrid fast PSO–BPN algorithm can be summarized as follows:

Algorithm:

Step 1: Initialize the positions and velocities of a group of particles randomly in the range of [0, 1].

Step 2: Evaluate each initialized particle's fitness value, and P_b is set as the positions of the current particles, while P_g is set as the best position of the initialized particles.

DATASET	TRAINING SET	TEST SET
SRBCT	63	20
LEUKEMIA	72	52
DLBCL	77	21

Step 3: If the maximal iterative generations is arrived, go to Step 8, else, go to Step 4.

Step 4: The best particle of the current particles is stored. The positions and velocities of all the particles are updated, then a group of new particles are generated, If a new particle files beyond the boundary [X_{min},X_{max}], the new position will be set as X_{min} or X_{max}, if a new velocity is beyond the boundary [V_{min},V_{max}], the new velocity will be set as V_{min} or V_{max}.

Step 5: Evaluate each new particle's fitness value, and the worst particle is replaced by the stored best particle. If the *i*th particle's new position is better than P_{ib}, P_{ib} is set as the new position of the ith particle. If the best position of all new particles is better than P_g, then P_g is updated.

Step 6: Reduce the inertia weights *w* according to the selection strategy described in PSO-BPN.

Step 7: If the current P_g is unchanged for ten generations, then go to Step 8; else, go to Step 3.

Step 8: Use the BP algorithm to search around P for some epochs, if the search result is better than P_g, output the current search result; or else, output P_g.

This algorithm has a parameter called learning rate that controls the convergence of the algorithm to an optimal local solution.

IV EXPERIMENTAL RESULTS

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

4.1 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 SRBCT dataset contains the expression data of 2308 genes. There are totally 63 training samples and 25 testing samples, 5 of the testing samples are not SRBCTs. The 63 training samples contain 23 Ewing family of tumors (EWS), 20 rhabdomyosarcoma (RMS), 12 Neuroblastoma (NB), and 8 Burkitt lymphomas (BL). And the 20 SRBCT testing samples contain 6 EWS, 5 RMS, 6 NB, and 3 BL.

B) Leukemia Dataset: The gene expression measurements were taken from 63 bone marrow samples and 9 peripheral blood samples. This dataset contains 72 samples. All samples can be divided into two subtypes: 25 samples of acute myeloid leukemia (AML) and 47 samples of acute lymphoblastic leukemia (ALL). The expression levels of 7129 genes were reported.

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 response to 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.

DATASET	NO OF GENE COMBINATION	ACCURACY (%)	
		Fast GA-BPN	Hybrid Fast PSO-BPN
LYMPHOMA	100, 2	67	82
LEUKEMIA	100, 3	71	85
SRBCT	100,4	76	89

Table 1 Summary of the Data Sets Used in the Experimentation

4.2 Testing Accuracy and Execution Time

In table 2 and 3 shows the accuracy and execution time for both BPN and Hybrid Fast GA-BPN are given in tabulation.

The table 2 represents the accuracy for gene expression data. The comparisons of fast GA-BPN and hybrid fast PSO-BPN approaches are evaluated using three datasets lymphoma, leukemia and SRBCT. The proposed method of hybrid fast PSO-BPN has high accuracy than fast GA-BPN.

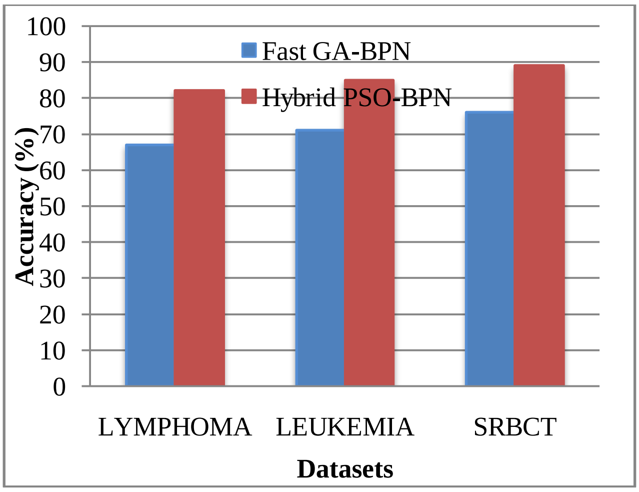

Fig. 2 shows the accuracy for Fast GA-BPN and Hybrid Fast PSO-BPN

Fig. 2 shows the accuracy value for fast GA-BPN and hybrid fast PSO-BPN. The proposed method of hybrid fast PSO-BPN has high accuracy when compare with other methods.

DATASET	NO OF GENES COMBINATION	EXECUTION TIME (Seconds)	
		Fast GA-BPN	Hybrid Fast PSO-BPN
LYMPHOMA	100, 2	47	40
LEUKEMIA	100, 3	34	29
SRBCT	100,4	25	19

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

Table 3 shows the execution time for GA-BPN and hybrid fast PSO-BPN. The execution time of the proposed hybrid approaches is less.

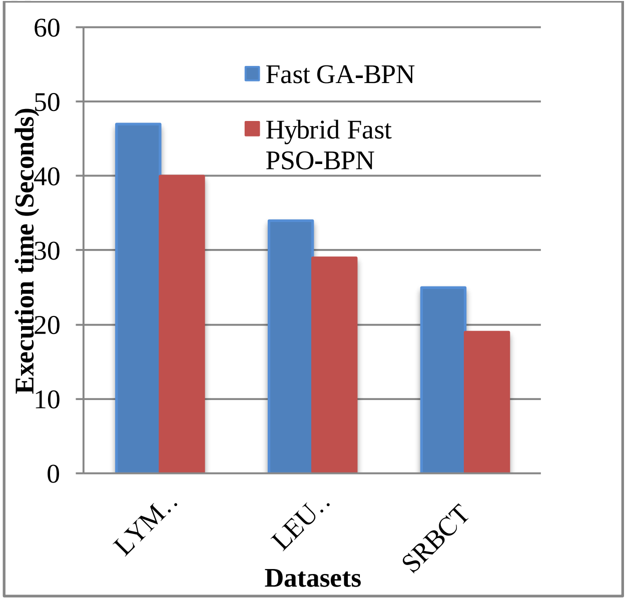

Fig. 3 shows the execution time of Fast GA-BPN and Hybrid Fast PSO-BPN

Fig. 3 shows the execution time for fast GA-BPN and hybrid fast PSO-BPN for gene expression data. The proposed method of hybrid fast PSO-BPN performs the function with less execution time. Hybrid fast PSO-BPN done better performance when compare with other.

V. CONCLUSION

A systematic and unbiased method for cancer classification is much significance to cancer treatment and drug discovery. Preceding classification methods are all clinical based and it does not used much because of less accuracy. The gene expression data classification using hybrid fast PSO-BPN may use list of datasets like Leukemia, Lymphoma and SRBCT for the proposed method. In this work, developed of a new hybrid method to solve a class of constrained global optimization problems. This method is based on the combination of the advantages of global exploration of fast GA-BPN and hybrid fast PSO-BPN. More precisely, the hybrid fast PSO-BPN technique was embedded into genetic algorithm as an acceleration operator during the iterations.

REFERENCES

1. Wang, Yu, Igor V. Tetko, Mark A. Hall, Eibe Frank, Axel Facius, Klaus FX Mayer, and Hans W. Mewes. "Gene selection from microarray data for cancer classification—a machine learning approach." *Computational biology and chemistry* 29, no. 1 (2005): 37-46.
2. Duan, Kai-Bo, Jagath C. Rajapakse, Haiying Wang, and Francisco Azuaje. "Multiple SVM-RFE for gene selection in cancer classification with expression data." *NanoBioscience, IEEE Transactions on* 4, no. 3 (2005): 228-234.
3. Statnikov, Alexander, Constantin F. Aliferis, Ioannis Tsamardinos, Douglas Hardin, and Shawn Levy. "A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis."*Bioinformatics* 21, no. 5 (2005): 631-643.

4. Zhu, Ji, and Trevor Hastie. "Classification of gene microarrays by penalized logistic regression." *Biostatistics* 5, no. 3 (2004): 427-443.
5. Weigelt, Britta, Frederick L. Baehner, and Jorge S. Reis-Filho. "The contribution of gene expression profiling to breast cancer classification, prognostication and prediction: a retrospective of the last decade." *The Journal of pathology* 220, no. 2 (2010): 263-280.
6. Peng, Sihua, Qianghua Xu, Xuefeng Bruce Ling, Xiaoning Peng, Wei Du, and Liangbiao Chen. "Molecular classification of cancer types from microarray data using the combination of genetic algorithms and support vector machines."*FEBS letters* 555, no. 2 (2003): 358-362.
7. Liu, Jane Jijun, Gene Cutler, Wuxiong Li, Zheng Pan, Sihua Peng, Tim Hoey, Liangbiao Chen, and Xuefeng Bruce Ling. "Multiclass cancer classification and biomarker discovery using GA-based algorithms." *Bioinformatics* 21, no. 11 (2005): 2691-2697.
8. Frohlich, Holger, Olivier Chapelle, and Bernhard Scholkopf. "Feature selection for support vector machines by means of genetic algorithm." In *Tools with Artificial Intelligence, 2003. Proceedings. 15th IEEE International Conference on*, pp. 142-148. IEEE, 2003.
9. Chen, X-W. "Gene selection for cancer classification using bootstrapped genetic algorithms and support vector machines." In *Bioinformatics Conference, 2003. CSB 2003. Proceedings of the 2003 IEEE*, pp. 504-505. IEEE, 2003.
10. Shah, Shital, and Andrew Kusiak. "Cancer gene search with data-mining and genetic algorithms." *Computers in Biology and Medicine* 37, no. 2 (2007): 251-261.
11. Chen, Kun-Huang, Kung-Jeng Wang, Min-Lung Tsai, Kung-Min Wang, Angelia Melani Adrian, Wei-Chung Cheng, Tzu-Sen Yang, Nai-Chia Teng, Kuo-Pin Tan, and Ku-Shang Chang. "Gene selection for cancer identification: a decision tree model empowered by particle swarm optimization algorithm." *BMC bioinformatics* 15, no. 1 (2014): 49.
12. Huang, Mei-Ling, Yung-Hsiang Hung, Wen-Ming Lee, R. K. Li, and Tzu-Hao Wang. "Usage of case-based reasoning, neural network and adaptive neuro-fuzzy inference system classification techniques in breast cancer dataset classification diagnosis." *Journal of medical systems* 36, no. 2 (2012): 407-414.
13. Sahab, M. G., A. F. Ashour, and V. V. Toropov. "A hybrid genetic algorithm for reinforced concrete flat slab buildings." *Computers & structures* 83, no. 8 (2005): 551-559.
14. Sadoughi, F., M. Ghaderzadeh, M. Solimany, and R. Fein. "An Intelligent System Based on Back Propagation Neural Network and Particle Swarm Optimization for Detection of Prostate Cancer from Benign Hyperplasia of Prostate." *J Health Med Informat* 5, no. 158 (2014): 2.
15. Khan, Koffka, and Ashok Sahai. "A comparison of BA, GA, PSO, BP and LM for training feed forward neural networks in e-learning context." *International Journal of Intelligent Systems and Applications (IJISA)* 4, no. 7 (2012): 23.