International Journal On Engineering Technology and Sciences – IJETS™
ISSN(P): 2349-3968, ISSN (O): 2349-3976
Volume III, Issue VI, June- 2016

A Novel Cancer Gene Search Model and Classification Using GPSO-BPNN

F.Leenavinmalar
M.Phil scholar,Department of computerscience,
chikkannagovt artscollege,Tirupur.
Dr.A.Kumarkombaiya
Assistant professor,Department of computer science,
chikkannagovt artscollege,Tirupur.

Abstract: Understanding the gene expression is a crucial issue to cancer designation. One target of this understanding is implementing cancer factor search and classification ways. However, cancer factor search and classification could be a challenge in this there's no an obvious precise algorithm which will be enforced severally for varied cancer cells. During this paper a pursuit is conducted through the info mining algorithms and enforced Geometric Particle Swarm optimization (GPSO) with Back propagation Neural Network (BPNN) for cancer factor search and classification, and the way they are enforced to achieve a higher performance. Hybrid GPSO-BPNN technique is employed to enhance the accuracy and higher convergence rate. This projected technique is used to beat from the matter of procedure difficulties occur by unwell-condition of the square penalty function. Finally, the experimental result shows that this projected technique is best in gene selection with less execution time.

Keywords:Cancer, Genes, Searching Algorithms, Classification Algorithms, Geometric Particle Swarm Optimization, Back propagation Neural Network

1. INTRODUCTION

In general, biological systems can be observed as information management system with a basic instruction set accumulated in each cell's DNA asgenes. The information for few genes is permitted when they are transcribed into RNA. Subsequently, which will betranslated into the proteinsthat structure much of a cell's mechanism and the phenomena is referred as gene expression [1]. The enormousbest part of fatal diseaseshas a unique gene expression profile which can be observed using microarray technology [2]. That gene profiling can throw in tobe cancer classification and profiling tumours. Gene expression profiling can permit the improvement of more suitable personalized treatment plans for individuals in view of the fact that some classes oftumours can be better treated with certain drugs than others, [3]. On the other hand, it is exceedingly complex to approach this manually withextensive numbers of genes requiring analysis. Consequently some of the techniques such as pattern recognition,statistical techniques and artificial intelligence are applied in DNAmicroarray research.

Usually, the classification of cancer withmicroarray data entails data acquisition and pre-processing, geneselection and classification [4]. Animportant initial step in microarray technology is data pre-processing which is done before data analysis is carried out [5]. Following pre-processing, the data can be signified in the form ofa matrixas represented in Fig. 1, where each row in the matrix corresponds toa particular gene where each column could either correspond to anexperimental condition or anexact time point at which expressionof the genes has been measured [6].

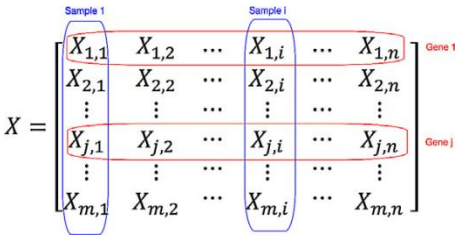

Fig. 1. Gene expression matrix

Applying statistical and procedure ways to microarray information could be a huge challenge, as this kind of information incorporates a high dimension. This can be oftentimes referred as 'the curse of dimension' within the literature [7]. In diseases like cancer, solely a number of genes are usually informative and need analysis [8]. The aim of factor choice is to pick those vital genes that contribute to

cancer and eliminate the remainder of the genes, in order that the dimension of the info is reduced for more investigation [9]. There are many shortcomings just the once the amount of genes is significantly over the amount of samples. As, each the interval and therefore the likelihood of misclassification area unit hyperbolic [10]. Once the genes are hand-picked, the classification procedure follows. During this procedure, first the classifier is trained so the classifier is employed to find the diagnostic class of latest samples [11]. There are several approaches which will be used for microarray organic phenomenon classification, like k nearest neighbours (k-NN) [12], Support Vector Machines (SVM) [13], Multilayer perceptron (MLP) [14], or alternative forms of Artificial Neural Networks (ANNs) [11].

The current work is to research the interval of every of the algorithms enforced so as to make a decision the simplest performance algorithm. So as to realize this goal, an innovative factor choice approach using GPSO-BPNN technique before cancer classification is projected. A completely unique optimization algorithm, GPSO is developed to boost classification performance.

2. RELATED WORK

Many works are done to prove that the Genetic algorithm plays a very important role in mining the fascinating patterns. Kulkarni et al. [15] mentioned the comparison between the accuracy of class prediction for 2 completely different classifiers particularly, Genetic programming and genetically evolved call tree. Jabbar et al. [16] projected an economical associative classification rule using genetic approach for the prediction of cardiopathy. Alshamla et al. [17] analyzed the performance of Bio inspired evolutionary gene selection methods. Anusha et al. [18] portrayed an increased K-means Genetic rule for optimum clustering. But there's a requirement for correct feature choice for higher additional optimum resolution [19].

Kabeer et al. [20] portrayed a hybrid approach of Boosted Feature set choice (BFSS) and Genetic rule (GA). Mourad et al. [21] used Genetic rule for mining the sequential patterns from the patient‟s prescription details using sequential interesting measures on Pharmacy information. Korayem et al. [22] have given a hybrid Genetic rule and artificial system for choosing gens from high dimensional DNA microarray dataset. Dipankar et al. [23] given a Multi Objective Genetic rule (MOGA) based mostly K-clustering methodology for optimizing the inter-cluster(separation, s)distance and intra-cluster (Homogeneity, H)distance. Marghny et al. [24] portrayed a good organic process agglomeration rule for the case study of viral hepatitis.

Peter et al. [25] mentioned and analyzed the effectiveness of Multi Objective k-means Genetic Algorithm (MOKGA). The additional work is required supported the classification of cancer factors in conjunction with another organic process methodology that would enhance the accuracy of gene patterns in medical dataset.

3. PROPOSED METHODOLOGY

In this work, we tend to have an interest in gene selection and classification of DNA Microarray information so as to differentiate tumor samples from traditional ones. For this purpose, we tend to propose hybrid model that use metaheuristics and classification techniques. The Particle Swarm optimization (PSO) combined with a BPNN approach. PSO could be a population based mostly metaheuristic impressed by the social behavior of bird flocking or fish schooling. Specifically, a recent version known as Geometric PSO has been employed in this work. The overall design diagram is given in fig.2.

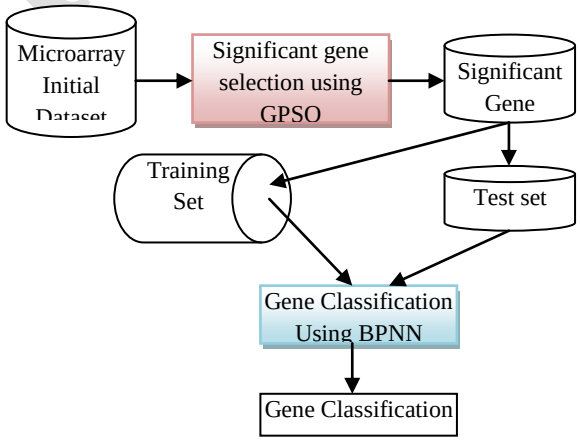

Fig.2.Architecture Diagram

Input Dataset

Data Sets Instances employed in this study consists of six wellknown datasets issued of microarray experiments such as ALLAML Leuke-mia dataset, carcinoma dataset, Colon tumor dataset, sex gland cancer dataset, prostate cancer dataset, and lung cancer dataset.

Geometric Particle Swarm Optimization

In this version, the situation of every particle i is depicted as vector $x_i = \langle x_{i1}, x_{i2}, \ldots, x_{iN} \rangle$Taking every bit x_{ij}(with j in {1, N}) binary values zero or one. The key issue of the GPSO is that the idea of particle movement. During this approach, rather than the notion of speed added to the position, a three-parent mask-based crossover (3PMBCX) operator is applied to every particle so as to "move" it. In line with the definition of 3PMBCX, given 3 parents a, b and c in $\hat{n}$, generate haphazardly a crossover mask of length n with symbols from the alphabet. Build the offspring filling every component with the bit from the parent showing within the crossover mask at the position.

The pseudocode of the GPSO formula for playing areas is illustrated in formula one. For a given particle (gene dataset) i, 3 parents participate within the 3PMBCX operator (line 13): this position ninety-one, the social best position user interface and therefore the historical best position found w_1, w_2And w_3(of this particle). The weight values w_1, w_2And w_3Indicate for every component within the crossover mask the chance of getting values from the fogeys ninety-one, x_i , g_i Or h_iSeverally. These weight prices associated to every parent represent the inertia value of this position w_1, the social influence of the global/local best position $w_2$andtherefore the individual influence of the historical best position foundw_3. A constriction of the geometric crossover forcesw_1, $w_2$and$w_3$to be non-negative and add up to at least one.

In summary, the GPSO developed during this study operates as follows: in a first part of the pseudocode, the format of particles are meted out by means that of the swarminitialization() perform (Line 1). This special format technique (used conjointly in our GA approach) was custom-made to cistron choice as follows. The swarm (population) was divided into four subsets of particles (chromosomes) initialized in several ways that reckoning on the amount of options in every particle. That is, 100 percent of particles were initialized with N (prefixed value) chosen genes (1s) set haphazardly.

Another two hundredth of particles was initialized with 2N genes, half-hour with 3N genes and at last, the remainders of particles (40%) were initialized haphazardly and five hundredth of the genes was turned on. In these experiments N are going to be adequate to four. In a second part, once the analysis of particles (line 4), historical and social position are updated (lines five to 10). Finally, particles are "moved" by means that of the 3PMBCX operator (line 13). Additionally, with a chance of 100 percent, an easy bit-mutation operator (line 14) is applied so as to avoid the first convergence. This method is recurrent till reach the stop condition fastened to a precise range of evolutions.

Algorithm 1:Pseudocode of the GPSO for Hamming space.

1: S ← Initializationswarm()
2: while not stop condition do
3: for each genex_iOf the swarm S do
4: evaluate(x_i)
5: if fitness(x_i) is better than fitness(h_i) then
6: $h_i \leftarrow x_i$
7: end if
8: if fitness(h_i) is better than fitness(g_i) then
9: $g_i \leftarrow h_i$
10: end if
11: end for
12: for each particle x_iOf the swarm S do
13: $x_i \leftarrow$ 3PMBCX((x_i, w_1), (g_i, w_2), (h_i, w_3))
14: mutate(x_i)
15: end for
16: end while
17: Output: Significant gene found

Evaluation Function

The entire fitness performs is represented as follows:

$$fitness(x) = \alpha \bullet (100/accuracy) + \beta \bullet \#genes,$$

Where α and β are weight values set to zero.75 and 0.25 severally so as to regulate that the accuracy price takes priority over the set size, since high accuracies are most well-liked once guiding the search method. The objectiveismaximizing the accuracy and minimizing the amount of genes. For convenience (only minimization of fitness) the primary issue is conferred as (100/accuracy).

Hybrid Geometric Particle Swarm improvement with Back propagation Neural Network

The procedure for this hybrid quick GPSO–BPNN formula is summarized as follows: Algorithm:

Step 1: Initialize the positions and velocities of a group of particles (genes) haphazardly within the value of [0, 1].

Step 2: measure every initialized particle's fitness price, and ninety-one is about because the positions of these particles, whereas user interface is about because the best position of the initialized particles.

Step 3: If the highest repetitive generations are arrived, attend Step eight, else, attend Step four.

Step 4: the simplest particle of this particles is hold on. The positions and velocities of all the particles are updated, then a group of latest particles are generated, If a brand new particle files on the far side the boundary [Xmin, Xmax], the new position are going to be set as Xmin or Xmax, if a brand new speed is on the far side the boundary [Vmin,Vmax], the new speed are going to be set as Vmin or Vmax. Rather than the notion of speed added to the position, a three-parent mask-based crossover (3PMBCX) operator is applied to every particle so as to "move" it.

Step 5: measure every new particle's fitness price, and therefore the worst particle is replaced by the hold on best particle. If the ith particle's new position is healthier than Pib, Pib is about because the new position of the ith particle. If the simplest position of all new particles is healthier than Pg, then Pg is updated.

Step 6: scale back the inertia weights w in line with the choice strategy represented in GPSO-BPNN.

Step 7: If this user interface is unchanged for 10 generations, then attend Step 8; else, attend Step three.

Step 8: Use the BPNN formula to go looking around P for a few epochs, if the search result's higher than user interface, output this search result; as an alternative, output g_i. This formula incorporates a parameter known as learning rate that controls the convergence of the formula to an optimum native resolution.

4. EXPERIMENTAL RESULTS AND DISCUSSION

Several observations can made based on the above experiments, so we tackle the analysis of results focusing on theperformance and robustness of our algorithms, as well asthe quality of the obtained solutions providing a biologicaldescription of most significant ones.

1.1. Accuracy Comparison

The proposed GPSO-BPNN method is having the average accuracy rate whereas the GA-BPNNmethod having lower accuracy results when compared to the proposed method. The overall accuracy percentage details are shown in fig 2.

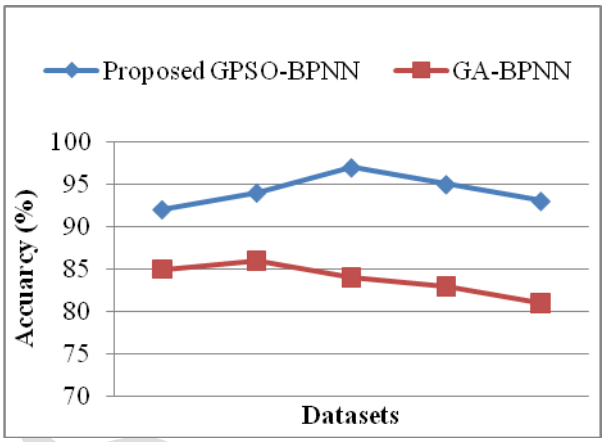

Fig.2.The Overall Accuracy Percentage Comparative Result

1.2. Precision Comparison

The Fig.3 shows the precision comparison result of existing GA-BPNN and proposed GPSO-BPNN algorithm. From the Fig.3, it is well known that the proposed system works better than existing system with the high precision result. The existing system has accuracy result which is less than the proposed GPSO-BPNN. The reason is that the proposed system has high convergence rate than the existing algorithms.

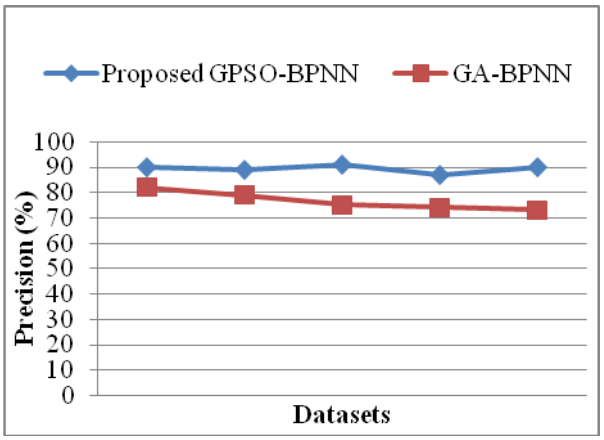

Fig.3.The Overall Precision Percentage Comparative Result

1.3. Recall Rate Comparison

The Fig.4 shows the recall comparison result of existing GA-BPNN and proposed GPSO-BPNN algorithm. From the Fig.4, it is obvious that the proposed system has high recall rate which is higher than the existing algorithms such as GA-BPNN. The reason is that the proposed system has less execution time than the existing algorithms.

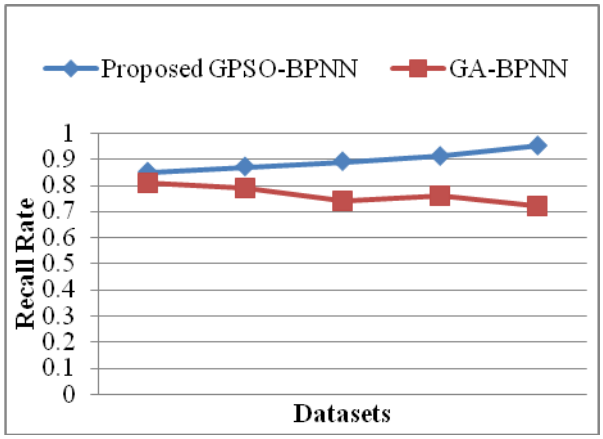

Fig.4.The Overall Recall Rate Comparative Result

5. CONCLUSION

In this work, two hybrid techniques for gene selection and classification of high dimensional DNA Microarray knowledge were given and compared. These techniques are supported completely different metaheuristic algorithms like GPSO used for feature selection using the BPNN classifier to spot probably sensible gene subsets. Specifically, the Geometric PSO rule for playacting area was wont to solve a true downside (gene selection during this case) for the primary time (to our knowledge). Additionally, genes elite are valid by a correct 10-fold cross validation methodology to enhance the particular classification. Results of 100% classification rate and few genes per set (3 and 4) are obtained in most of our executions. Continued the road of this work, we have an interest in developing and testing many combos of alternative metaheuristics with classification strategies so as to find new and higher subsets of genes using specific Microarray datasets.

References

1. C. Ambroise, G. Mclachlan, Selection bias in gene extraction on the basis of microarray gene-expression data, Proc. Natl. Acad. Sci. 99 (10) (2002) 6562–6566.
2. R. Breitling, P. Armengaud, A. Amtmann, P. Herzyk, Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments, FEBS Lett. 573 (1–3) (2004) 83–92.
3. T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, et al., Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, Science 286 (5439) (1999) 531–538.
4. A. Rahideh, M.H. Shaheed, Cancer classification using clustering based gene selection and artificial neural networks, in: International Conference on Control, Instrumentation and Automation (ICCIA), Shiraz, 2011, pp. 1175–1180.
5. P. Baldi, G.W. Hatfield, DNA Microarrays and Gene Expression: From Experiments to Data Analysis and Modeling, Cambridge University Press, Cambridge, 2002.
6. B.M. Madan, Introduction to microarray data analysis, in: Grant (Ed.), Computational Genomics, Horizon Press, 2004.
7. R.L. Somorjai, B. Dolenko, R. Baumgartner, Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions, Bioinformatics 19 (12) (2003) 1484–1491.
8. W. Xiong, Z. Cai, J. Ma, ADSRPCL-SVM approach to informative gene analysis, Genomics Proteomics Bioinform. 6 (2) (2008) 83–90.
9. M.S. Mohamad, S. Omatu, S. Deris, M. Yoshioka, A modified binary particle swarm optimization for selecting the small subset of informative genes from gene expression data, IEEE Trans. Inf. Technol. Biomed. 15 (6) (2011) 813–822.
10. Y. Saeys, I. Inza, P. Larranaga, A review of feature selection techniques in bioin- formatics, Bioinformatics 23 (19) (2007) 2507–2517.
11. M.H. Asyali, Gene expression profile classification: a review, Curr. Bioinform. 1 (1) (2006) 55–73.
12. C. Li, S. Zhang, H. Zhang, L. Pang, et al., Using the K-nearest neighbor algorithm for the classification of lymph node metastasis in gastric cancer, Comput. Math. Methods Med. 2012 (2012).
13. T.S. Furey, N. Cristianini, N. Duffy, D.W. Bednarski, et al., Support vector machine classification and validation of cancer tissue samples using microarray expression data, Bioinformatics 16 (10) (2000) 906–914.
14. Z.Y. Wang, Y. Wang, J.H. Xuan, Y.B. Dong, M. Bakay, et al., Optimized multilayer perceptrons for molecular classification and diagnosis using genomic data, Bioinformatics 22 (6) (2006) 755–761.

15. AshwinikumarKulkarni, B.S.C. Naveen Kumar, Vadlamani Ravi, UpadhyayulaSuryanarayana Murthy, "Colon Cancer Prediction with Genetic profiles using evolutionary techniques", Elsevier, pp.2752-2757,2011.
16. M.Akhiljabbar, Dr. Priti Chandra and Dr. B.L Deekshatulu, "Heart Disease Prediction System using Associative Classification ang Genetic algorithm", International Conference on Emerging Trends in Electrical, Electronics and Communication Technologies, 2012.
17. Hala M. Alshamlan, Ghada H. Badr and Yousef A. Alohali, "The performance of Bio-Inspired Evolutionary Gene Selection Methods for Cancer Classification using Microarray Dataset", International Journal of Bioscience, Biochemistry and Bioinformatics, pp.166- 170, 2014.
18. M.Anusha and J.G.R.Sathiaseelan, "An Enhanced K-means Genetic Algorithms for Optimal Clustering", IEEE ICCIC, pp.580-584, 2014.
19. M.Anusha and J.G.R.Sathiaseelan, "An Improved K-Means Genetic Algorithm for Multi-objective Optimization", International Journal of Applied Engineering Research, pp. 228-231, 2015.
20. ShaikhJeeshanKabeer, MoinMhmudTanvee, Mohammad ArifurRahman, Abdul Mottalib, Md. HasanumKabir, "BFFS: Enhancing the performance of Genetic Algorithm using Boosted Filtering Approach", International Journal of Computer Applications, pp.29-34, 2012.
21. MouradYkhlef and Hebahelgibreen, "Mining Pharmacy Database Using Evolutionary Genetic Algorithm", International Journal of Electronics and Telecommunications, pp. 427-432, 2010.
22. Mohammed Korayem, Waleed Abo Hamad, KhaledMostafa, " A Hybrid Genetic Algorithm and Artificial immunse system for informative gene selection", International Journal of Computer Science and Network Security, pp.76-83, 2010.
23. DipankarDutta, ParamarthaDutta, Jaya Sil, "Data clustering with mixed features by Multi Objective Genetic Algorithm", IEEE, pp.336-341, 2012.
24. M.H. Marghny, Rasha M. Abd El-Aziz, Ahmed I. Taloba, "An effective evolutionary clustering algorithm: Hepatits C case study", International Journal of Computer Applications, pp.1-6, 2011.
25. Peter Peng, Omer Addam, MohamadElzohbi, Sibel T. Ozyer, Ahmad Elhajj, Shang Gao, Yimin Liu, TanselOzyer, Mehmet Kaya, Mick Ridley, Jon Rokne, RedaAlhajj, "Reporting and analyzing alternative clustering solutions by employing multi-objective genetic algorithm and conducting experiments on cancer data", Elsevier, pp.108-122, 2014.