Asian Journal of Information Technology 13 (12): 761-769, 2014
ISSN: 1682-3915

Cooperative Parallel Multi-Objective Genetic Algorithm for Gene Feature Selection to Diagnose Breast Cancer

[1]A. Natarajan and [2]T. Ravi
[1]Department of Information Technology,
Jayaraj Annapackiam CSI College of Engineering, Nazareth, Tamilnadu, India
[2]Srinivasa College of Engineering and Technology, Chennai, India

Abstract: Gene selection is very important in classification of cancer using parallel computing in the analysis of gene expression relationship. The high performance parallel computing is used for gene expression analysis and finding the thousands of genes simultaneously. DNA microarrays are used to measure the expression levels of thousands of genes simultaneously. The classification and validation of molecular biomarkers for cancer diagnosis is an important problem in cancer genomics. The microarray data analysis is very much important to extract biologically useful data from the huge amount of expression data to know the current state of the cell. Most cellular processes are regulated by changes in gene expression. This is a great challenge for computational biologists who see in this new technology the opportunity to discover interactions between genes. In this study, we propose a Cooperative Parallel Multi-Objective Genetic algorithm for Gene Feature Selection. We have implemented CPMGA for gene feature selection to classify the breast cancer data sets. More importantly, the method can exhibit the inherent classification difficulty with respect to different gene expression datasets, indicating the inherent biology of specific cancers.

Key words: Microarray, gene expression, Multi-Objective Genetic algorithm, gene feature selection, Island model, parallel GA

INTRODUCTION

DNA microarray experiments play a very important role in cancer classification and prediction. The microarray technology has been used in many cancer researches. The analysis of the large volumes of data generated under different experimental conditions is important and it requires advanced knowledge discovery methods. Many data mining techniques (Witten and Frank, 2005) have been proposed to analyze microarray data (Wang and Gotoh, 2010). Feature Selection (FS) is a very important task in data mining for cancer classification with the goal of identifying very important features subsets in a microarray data. It is one of the most key problems in the field of machine learning. The classification and validation of molecular biomarkers for cancer diagnosis is an important problem in cancer genomics. The selection of candidate genes is very crucial to identify accurately the origin of cancer, its treatment and diagnosis too. With, the appearance and fast development of DNA microarray technologies, making gene expression profiles for different cancer types has already become a hopeful means for cancer classification.

Genetic Algorithms (GAs) (Goldberg and Holland, 1988), a form of inductive learning strategies are adaptive search techniques initially introduced by Holland. Genetic algorithms are inspired from Darwin's theory of evolution. By simulating nature evolution and emulating biological selection and reproduction techniques, the GA can solve complex problems in a strong search domain. The algorithm starts with a set of randomly generated solutions called population. The population size remains constant throughout the genetic algorithm. At each iteration the populations are evaluated based on their fitness quality with respect to the given application domain to form new solutions called offspring which retains many features of their parents. Offsprings are formed by two main Genetic algorithm operators such as crossover and mutation. Crossover operates by randomly selecting a point in the two selected parent gene

Corresponding Author: A. Natarajan, Department of Information Technology, Jayaraj Annapackiam CSI College of Engineering, Nazareth, Tamilnadu, India

structures and exchanging the remaining segments of the parents to create new offspring. Therefore, crossover combines the features of two individuals to create two similar offsprings. Mutation operates by randomly changing one or more components of a selected individual. It acts as a population perturbation operator and is a means for inserting new information into the population. This operator prevents any stagnation that might occur during the search process.

In this proposed research, CPMOGA feature selection is implemented based on Multi-Objective Genetic algorithm. CPMOGA uses a different operator called multi-objective operator. Multi-objective aspect is defined to find the pareto optimal solutions for ranking. Since, the search space is large and requires a good diversity, Island model has been proposed. Finally, the cooperative parallel GA has been implemented by using parallelization tools (Umbarkar and Joshi, 2013) (Open MP).

Literature review: Pati *et al.* (2013) has explained a novel feature selection method which was based on Multi-Objective Genetic algorithm using rough set theory. This method proposed to choose important informative gene set which classify the cancer dataset very efficiently. This method has used two fitness functions individually based on the concepts of strong mathematics such as rough set theory and probability theory. The lack of diversity of population is overcome by jumping gene mutation. The only drawback of this method is that the population size can be set within the range 100-1000 only.

Karegowda *et al.* (2010) has proposed a wrapper approach with Genetic algorithm for generation of subset of attributes with different classifiers such as naive bayes, bayes networks, C4.5 and radial basis functions. The above classifiers are experimented on the diabetes datasets, breast cancer datasets, heart statlog and wisconsin breast cancer. The main disadvantage of this approach is that the computing time is very high for the large datasets.

Liu *et al.* (2009) has proposed a new feature selection method called recursive feature addition method on microarray based breast cancer data. The *RFA* gene feature selection method provides good classification accuracy than the other methods. In this method, serial programming is used for classification which slows down the computational speed.

MATERIALS AND METHODS

Genetic algorithm operators for generating population using Island model: The proposed GA based cooperative parallel multi-objective GA is implemented for selecting the most important genes from the breast cancer data set. The population generation is defined in GA by two operators (Khabzaoui *et al.*, 2006) crossover and mutation. The cross over operator has two versions Crossover by exchanging values crossover by inserting values. The crossover operator mixes the features of two rules by the combination of their attributes. In the proposed research, crossover operators can be defined as:

Crossover by exchanging values: If two rules X and Y have one or several common attribute(s) in their C parts, one common attribute is randomly selected. The value of the selected attribute in X is exchanged with its counterpart in Y (Fig. 1).

Crossover by inserting values: Conversely, if X and Y have no common attribute, one term is randomly selected in the C part of X and inserted in Y with a probability inversely proportional to the length of Y. The related operation has performed to insert one term of Y in X (Fig. 2).

Mutation is a genetic algorithm operator. This operator changes at random the value of a gene in a newly created individual. The population mutation rate used is very small, in the order of one mutation per thousand genes transfer. Thus, mutation is considered to be a secondary mechanism of genetic algorithms. It is still used to introduce new solutions into the population and to protect the algorithm from premature loss of important genetic material by reintroduction of genes. Genetic algorithm has four mutation operators value mutation operator attribute mutation insertion operator delete operator. In this study, we proposed adaptive strategy for calculating the rate of application of each mutation. We compute the new rate of mutation by calculating the progress of the jth application of mutation M_i for an individual ind mutated into an individual mut as follows:

$$\text{Progress}_j(M_i) = \text{Max}(\text{fitness}(\text{ind}, \text{fitness}(\text{mut})) - \text{fitness}(\text{ind})) \tag{1}$$

Then, for each mutation operator M_i, assume Nb_mut(M_i) applications of the mutation are done during a given generation (j = 1, ..., Nb_mut(M_i)). Then, we can compute the profit of a mutation M_k:

$$\text{Profit}(M_k) = \frac{\sum_j \text{Progress}_j(M_k)/\text{Nb\_mut}(M_k)}{\sum_i\left(\sum_j \text{Progress}_j(M_i)/\text{Nb\_mut}(M_i)\right)} \tag{2}$$

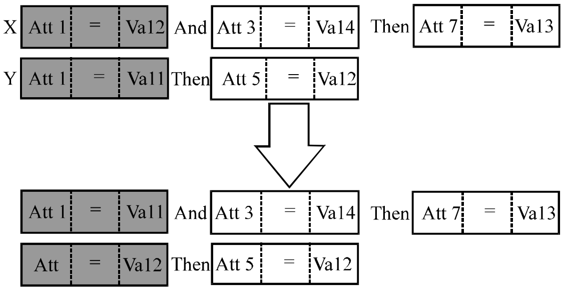
X Att 1 = Va12 And Att 3 = Va14 Then Att 7 = Va13
Y Att 1 = Va11 Then Att 5 = Va12
Att 1 = Va11 And Att 3 = Va14 Then Att 7 = Va13
Att = Va12 Then Att 5 = Va12

Fig. 1: Crossover by exchanging value

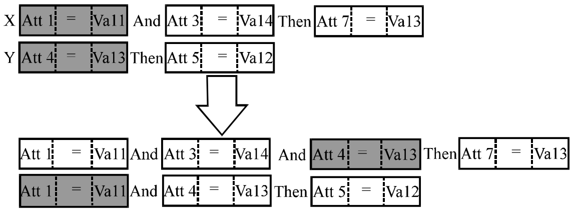
X Att 1 = Va11 And Att 3 = Va14 Then Att 7 = Va13
Y Att 4 = Va13 Then Att 5 = Va12
Att 1 = Va11 And Att 3 = Va14 And Att 4 = Va13 Then Att 7 = Va13
Att 1 = Va11 And Att 4 = Va13 Then Att 5 = Va12

Fig. 2: Crossover by insertion of attributes

We set a minimum rate δ and a global mutation rate $p_{mutation}$ for N mutation operators. The new mutation ratio for each M_i is calculated using Eq. 3 (Hong *et al*., 2000):

$$p(M_i) = Profit(M_i)\times(p_{mutation}-N\times\delta)+\delta \quad (3)$$

The sum of all the mutations is equal to the global rate of mutation $p_{mutation}$. The initial rate of application of each mutation operator is set to $p_{mutation}/N$.

Multi-Objective Genetic algorithm: In the Single Objective Genetic algorithm , the classification accuracy is low and the time taken is very high using serial computing. To overcome this problem researchers proposed cooperative parallel Multi-Objective Genetic algorithm that has been implemented in parallel computing. The population generation is one of basic step in Multi-Objective Genetic algorithm applied here for dimension reduction. The population generation is derived by using the multiobjective optimization problem (Van Veldhuizen and Lamont, 2000). In a multi-objective optimization, all the solutions are best compromise. The best solutions encountered over generations are filed into a secondary population called the Pareto archive and the solutions can be selected from the Pareto archive during the production process. This method is called as elitism. The offspring solutions replace their parents according to the replacement strategy. Figure 3 presents the Multi-Objective Genetic algorithm scheme.

Cooperative parallel Multi-Objective Genetic algorithm: The population generation is one of the basic steps in cooperative Multi-Objective Genetic algorithm. The Parallel Genetic algorithm (Konfrst, 2004) has been classified into three main models: global, fine cellular and island. The global model uses parallelism to speed up the sequential GA. This model uses a global shared population and the fitness evaluation is done on different processors. The cellular model seeks to exploit the fine-gained, massively parallel architectures. The population is separated into a large number of very small sub-populations which are maintained by different processors. In the island model, the population is divided into a few large independent subpopulations called islands. Each processor evolves their population using a parallel GA. For each island, some solutions rarely migrate to another island. We choose the island model for Parallel CPMOGA.

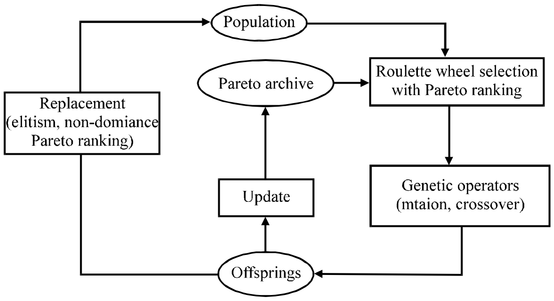
Population
Pareto archive
Roulette wheel selection with Pareto ranking
Replacement (elitism, non-domiance Pareto ranking)
Update
Genetic operators (mtaion, crossover)
Offsprings

Fig. 3: A multi-objective GA

Island model: The island model is implemented in parallel programming and many islands are connected by using ring topology (Fig. 4). This model typically runs on a parallel multi-objective GA. In this, each processor is called an island with independent populations and Pareto archives (Fig. 5). Each GA starts with its proper parameters such as population, parameters of GA. Periodically, each Island sends some solutions from its Pareto archive randomly selected to the neighboring Island. The Island model has four steps:

- Each Island model creates its population
- Each model develops its population for a global number of generations and updates its archives for every generation
- Each island sends some solutions of its Pareto archive to the neighboring island according to the migration policy
- Finally, the island receives all the migrating solutions and replaces its worst solutions by those immigrants according to the ranking

At the end, a specific Island waits for all the others to finish their execution and collects all the final Pareto archives to create the global Pareto archive. This island model has been implemented by using parallelization tools like Open MPI (August *et al*., 2010; Quinn, 2004).

Code design and selecting best parameters: In the code design the default values used are for selecting the best parameters.

- Population size is 1000
- Selection in population is 4/6 (400)
- Global mutation rate is 0.5
- Crossover rate is 0.8
- Selection in Pareto archive (elitism) is 0.5
- Minimal number of generations are 500

The stopping criterion used is the nonamelioration. Once the minimal number of generations has been overpassed and after taking the best solution for all 10 generations, the iteration stops. For further improving the computing time, we have used clustered machine comprised of six workstations with Intel Core i5 processor with NVIDIA Graphics processor (Oiso *et al*., 2011) and 1 GB main memory. The study shows the pseudo code for the parallel Island model:

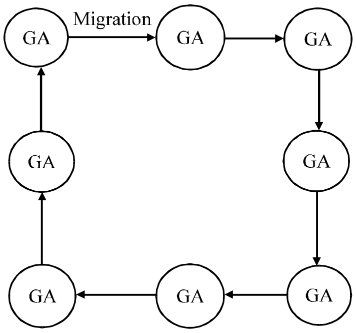

Fig. 4: Island model ring topology connection

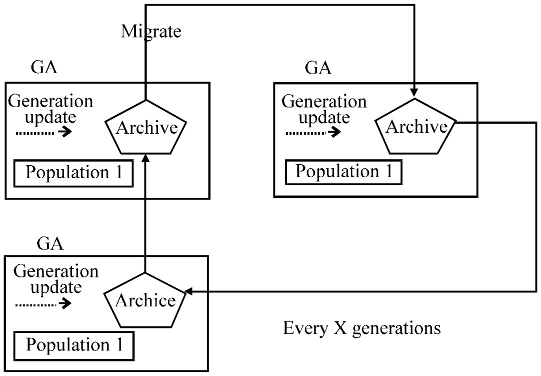

Fig. 5: Island model

Island model pseudo code:

```
Island Model (A, n, $\mu$)
Begin
Concurrently for each of the i? 1 to n subpopulations
initialize ($P_i$, $\mu$)
For each no of generation? 1 to A do
Concurrently for each of the i? 1 to n subpopulations do
Sequential_GA($P_i$, $G_i$)
Od;
For i? 1 to n do
For each neighbour j of i
Migration ($P_i$, $P_j$).
Assimilate ($P_i$);
od
od
Problem solution = best individual of all subpopulations;
End
```

The sequential GA has implemented using the following pseudo code:

Sequential GA pseudo code:

```
Sequential_GA ($P_i$, $P_j$)
Begin
For generation? 1 to $G_i$ do
Pnew? P?
For offspring ? 1 to Max_offspring do
$P_\alpha$? selection($P_i$)
$P_\beta$? selection($P_i$)
Pnew = pnew? Crossover ($P_\alpha$, $P_\beta$)
Od
Fitness_calculation ($P_i$?$P_{new}$)
$P_i$? Reduction ($P_i$?$P_{new}$)
Mutation ($P_i$);
Fitness_calculation ($P_i$)
End
```

ROC curve of breast cancer data sets for data mining classifiers (Fig. 6-11).

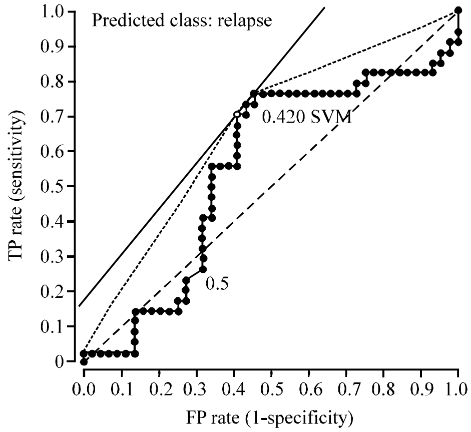

Fig. 6: ROC analysis for SVM classifier; classifier: SVM; target class: relapse; costs: FP = 500, FN = 500; prior target class probability: 43%

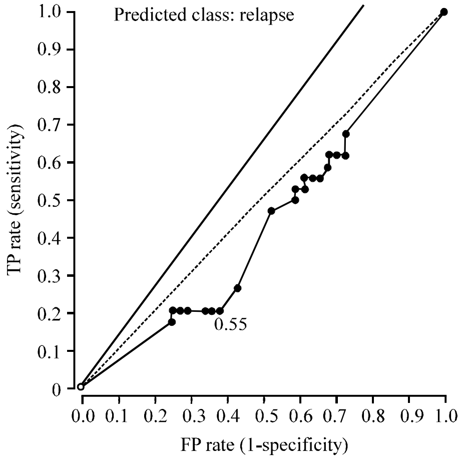

Fig. 7: ROC analyses for CN2 classifier; classifier: CN2 rules; target class: relapse; costs: FP = 500, FN = 500; prior target class probability: 43%

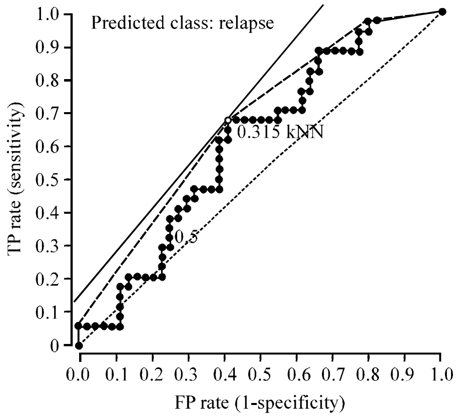

Fig. 8: ROC analysis for kNN classifier; classifier: kNN; target class: relapse; costs: FP = 500, FN = 500; prior target class probability: 43%

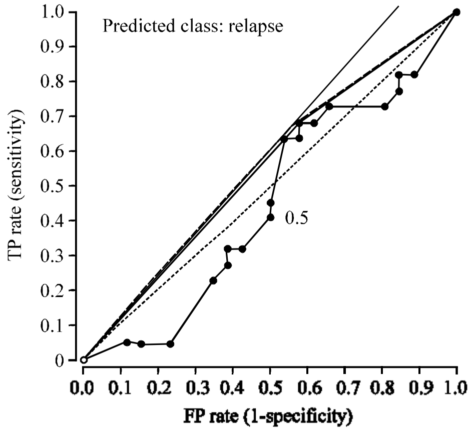

Fig. 9: ROC analysis for classification tree; classifier: tree; target class: yes; costs: FP = 500, FN = 500; prior target class probability: 45%

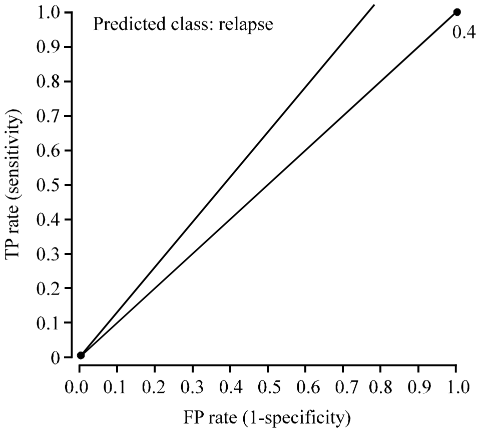

Fig. 10: ROC analysis for ITB rules classifier; classifier: intractive tree builder; target class: relapse; costs: FP = 500, FN = 500; prior target class probability: 43%

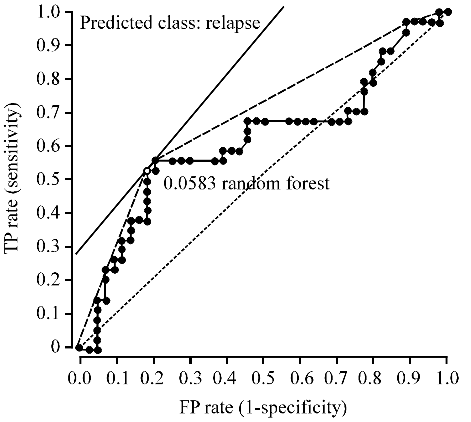

Fig. 11: ROC analysis for random forest classifier; classifier: rendom forest; target class: relapse; costs: FP = 500, FN = 500; prior target class probability: 43%

RESULTS AND DISCUSSION

The proposed method has been implemented in Multicore processor environment. The measured run times for 4, 8, 16, 32 processors are shown in the Table 1. The best features have been taken from *CPMOGA* gene Feature Selection Method and compared with the existing feature selection methods using the orange data mining tools (http://orange.biolab.si/docs/latest/tutorial/rst/) using python scripting language.

We carried out this experiment on two publicly available microarray breast cancer datasets (Mendes, 2011) available at Kent Ridge Bio-medical Data Set Repository (http://datam.i2r.a-star.edu.sg/datasets/krbd/). The existing feature selection methods are applied for this breast cancer data sets and classification accuracy are measured using Orange data mining and machine Learning tool. The feature selection methods (Hassanien, 2003; Polat *et al.*, 2005) like Novel Hybrid, Wrapper approach, Consistency Subset Selection (CON) (Yu and Liu, 2004), Correlated Feature Selection (CFS) (Hall, 1999), Single Objective Genetic Algorithm (SOGA) (Goldberg and Holland, 1988), Multi-Objective GA (MOGA) and the proposed Cooperative Parallel Multi Objective GA (CPMOGA) are applied on breast cancer data sets. It is then measured by various classification methods like random forest, K-NN, classification tree, SVM, CN2 rules and interactive tree builder in orange tool. The proposed CPMOGA Feature Selection Method provides better classification accuracy than the other Feature Selection algorithm and the computing time is very less than the other methods. Table 2 shows the classification accuracy of various methods. The Cross-fold validation is used for measuring the classification accuracy in terms of time which is available at Orange tool and CPMOGA is implemented by us. From Table 2 researchers observe that the proposed CPMOGA method is better than other methods in terms of time consuming, feature selection and classification accuracy. The executing system consists of core i5 processor with NVIDIA Graphics processor running at 4 GHz clock frequency.

Table 1: Measured run time

No. of processors	Run time (h)
4	3.50
8	1.78
16	0.90
32	0.45

Some statistical measurements like true positive rate, false positive rate of the classifiers are calculated using Eq. 4 and 5:

$$\text{TPR} = \frac{\text{TP}}{\text{P}} = \frac{\text{TP}}{(\text{TP} + \text{FN})} \tag{4}$$

$$\text{FPR} = \frac{\text{FP}}{\text{N}} = \frac{\text{FP}}{(\text{P} + \text{TN})} \tag{5}$$

Table 2: Classification accuracy (%) of breast cancer data set sub-types

		Data mining classifier					
Breast cancer data set sub-types	Feature selection methods	SVM	ITB	KNN	CT	CN2	RF
Normal breast like	Wrapper Approcah	62.0	54.0	58.0	54.0	45.0	51
	CON	61.0	53.0	59.0	55.0	48.0	53
	CFS	60.0	53.0	60.0	57.0	48.0	51
	SOGA	62.0	51.0	61.0	56.0	48.0	50
	MOGA	69.0	59.0	63.0	55.0	49.0	48
	CPMOGA	76.0	72.0	72.0	78.0	50.5	54
	Novel Hybrid	65.0	55.0	65.0	89.0	50.0	51
Basal	Wrapper Approcah	62.0	52.0	55.0	53.0	43.0	50
	CON	60.0	52.0	53.0	53.0	48.0	53
	CFS	60.0	53.0	60.0	57.0	48.0	51
	SOGA	69.0	65.0	67.0	58.0	54.0	66
	MOGA	72.0	74.0	75.0	71.0	64.0	68
	CPMOGA	73.0	74.5	76.0	76.0	69.0	72
	Novel Hybrid	61.0	61.0	69.0	71.0	64.0	68
Luminal A	Wrapper Approcah	64.0	65.0	61.0	68.0	46.0	56
	CON	69.0	66.0	61.0	53.0	46.0	57
	CFS	69.0	65.0	67.0	58.0	54.0	66
	SOGA	73.0	67.0	67.0	65.0	56.0	66
	MOGA	76.0	73.0	71.0	73.0	55.0	57
	CPMOGA	76.5	73.5	72.0	76.0	67.0	68
	Novel Hybrid	74.0	75.0	74.0	73.0	63.0	61
Luminal B	Wrapper Approcah	74.0	76.0	73.0	72.0	56.0	59
	CON	73.0	74.0	72.0	70.0	57.0	55
	CFS	70.0	72.0	74.0	70.0	54.0	51
	SOGA	68.0	69.0	67.0	68.0	52.0	57
	MOGA	72.0	73.0	70.0	68.0	66.0	67
	CPMOGA	74.0	75.0	71.0	73.0	68.0	70
	Novel Hybrid	69.0	70.0	70.0	71.0	65.0	65
HER2+/ER	Wrapper Approcah	65.0	66.0	63.0	53.0	51.0	59
	CON	63.0	64.0	64.0	55.0	55.0	57
	CFS	64.0	65.0	66.0	56.0	57.0	59
	SOGA	67.0	70.0	71.0	69.0	58.0	61
	MOGA	70.0	71.0	72.0	67.0	59.0	62
	CPMOGA	73.0	74.0	73.5	70.5	61.0	62
	Novel Hybrid	70.0	71.0	68.0	69.0	58.0	59

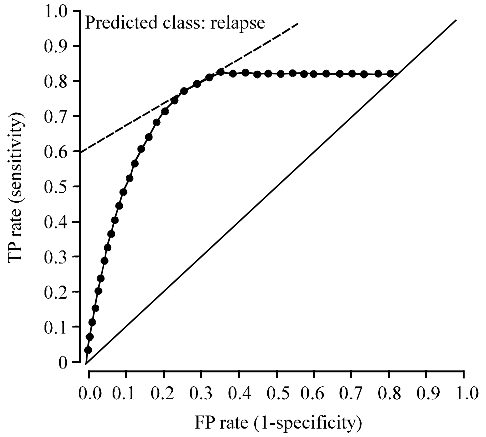

Fig. 12: Proposed ROC for SVM; classifier: SVM; target class: relapse; costs: FP = 500, FN = 500; prior target class probability: 53%

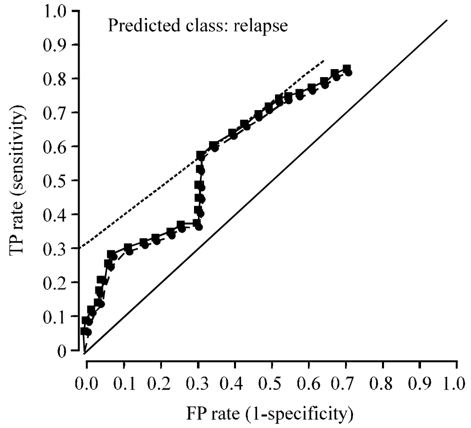

Fig. 13: Proposed ROC for RF; classifier: random forest; target class: relapse; costs: FP = 500, FN = 500; prior target class probability: 53%

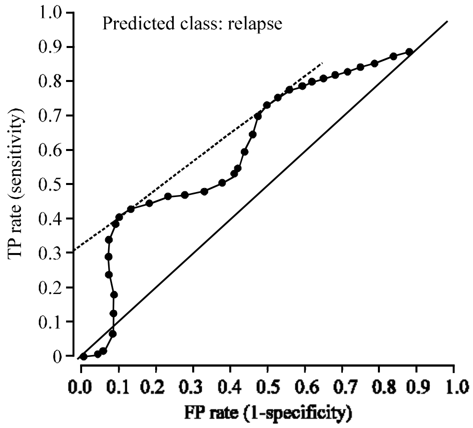

Fig. 14: Proposed ROC for kNN; classifier: kNN; target class: relapse; costs: FP = 500, FN = 500; prior target class probability: 53%

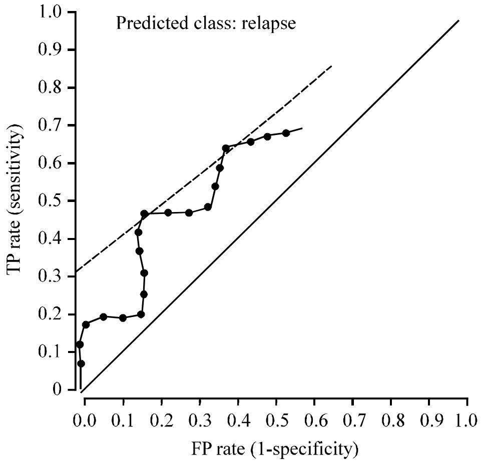

Fig. 15: Proposed ROC for CT; classifier: classification tree; target class: relapse; costs: FP = 500, FN = 500; prior target class probability: 53%

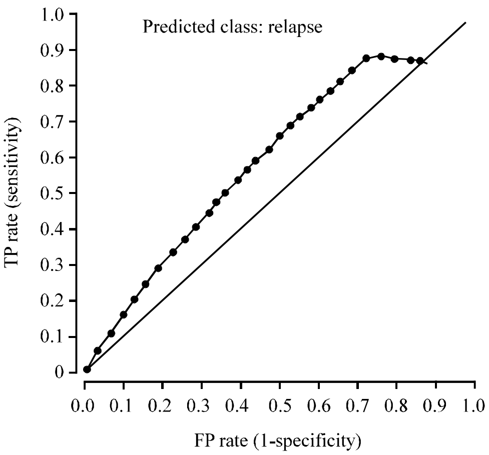

Fig. 16: Proposed ROC for ITB; classifier: inter active tree builder; Target class: relapse; costs: FP = 500, FN = 500; prior target class probability: 53%

Where:

TP = Positive object classified as positive
FP = Positive object classified as negative
TN = Negative object classified as negative
FN = Negative object classified as positive

The ROC curves visualized by orange data mining tool for the existing feature selection methods for different classifiers. It shows the classification performance of the different classifier. The proposed ROC curve is shown in Fig. 12-17. From the graph, we observed that the true positive rate is better than the existing methods. The proposed method curves

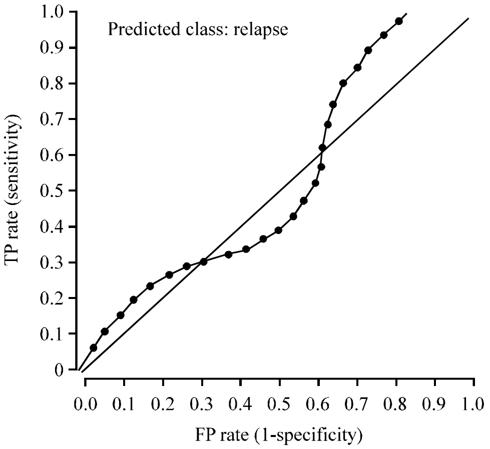

Fig. 17: Proposed ROC for CN2; classifier: CN2; target class: relapse; costs: FP = 500, FN = 500; prior target class probability: 53%

indicates that the best features are taken and provides accurate classification performance on the breast cancer data set.

CONCLUSION

In the above research, the cooperative parallel Multi-Objective Genetic algorithm has been implemented and best features are selected in short time. The gene feature selection is very important in cancer classification. This method uses the island model for generating the best population. The multiple islands are implemented in parallel which has substantially reduced the execution time in the process of best feature selection. In this research, standard microarray data sets are taken from kent ridge bio-medical data set repository. In future real time data from breast cancer patients has to be taken. The classification accuracy should also be clinically verified.

REFERENCES

August, A.D., K.P.D. Chiou, R. Sendag and J.J. Yi, 2010. Programming multicores: Do applications programmers need to write explicitly parallel programs? IEEE Micro, 30: 19-33.

Goldberg, D.E. and J.H. Holland, 1988. Genetic algorithms and machine learning. Mach. Learn., 3: 95-99.

Hall, M.A., 1999. Correlation-Based Feature Selection for Machine Learning. University of Waikato Press, New Zealand, Pages: 178.

Hassanien, A.E., 2003. Classification and feature selection of breast cancer data based on decision tree algorithm. Stud. Inform. Control, 12: 33-39.

Hong, T.P., H.S. Wang and W.C. Chen, 2000. Simultaneously applying multiple mutation operators in genetic algorithms. J. Heurist., 6: 439-455.

Karegowda, A.G., M.A. Jayaram and A.S. Manjunath, 2010. Feature subset selection problem using wrapper approach in supervised learning. Int. J. Comp. Applic., 1: 13-17.

Khabzaoui, M., C. Dhaenens and E.G. Talbi, 2006. A Cooperative Genetic Algorithm for Knowledge Discovery in Microarray Experiments. In: Parallel Computing for Bioinformatics and Computational Biology, Zomaya, A.Y. (Ed.). John Wiley and Sons, New York, ISBN-13: 9780471756491, pp: 303-324.

Konfrst, Z., 2004. Parallel genetic algorithms: Advances, computing trends, applications and perspectives. Proceedings of the 18th International Parallel and Distributed Processing Symposium, April 26-30, 2004, USA.

Liu, Q., A.H. Sung, Z. Chen, J. Liu, X. Huang and Y. Deng, 2009. Feature selection and classification of MAQC-II breast cancer and multiple myeloma microarray gene expression data. PloS One, Vol. 4. 10.1371/journal.pone.0008250.

Mendes, A., 2011. Identification of breast cancer subtypes using multiple gene expression microarray datasets. Proceedings of the 24th Australasian Joint Conference on Advances in Artificial Intelligence, December 5-8, 2011, Perth, Australia, pp: 92-101.

Oiso, M., T. Yasuda, K. Ohkura and Y. Matumura, 2011. Accelerating steady-state genetic algorithms based on CUDA architecture. Proceedings of the IEEE Congress on Evolutionary Computation, June 5-8, 2011, New Orleans, LA., USA., pp: 687-692.

Pati, S.K., A.K. Das and A. Ghosh, 2013. Gene selection using multi-objective genetic algorithm integrating cellular automata and rough set theory. Proceedings of the 4th International Conference on Swarm, Evolutionary and Memetic Computing, December 19-21, 2013, Chennai, India, pp: 144-155.

Polat, K., S. Sahan, H. Kodaz and S. Gunes, 2005. A new classification method for breast cancer diagnosis: Feature Selection Artificial Immune Recognition System (FS-AIRS). Proceedings of the 1st International Conference on Advances in Natural Computation, August 27-29, 2005, Changsha, China, pp: 830-838.

Quinn, M.J., 2004. Parallel Programming in C with MPI and OpenMP. Tsinghua University Press, Beijing, China, ISBN-13: 9787302111573, Pages: 519.

Umbarkar, A.J. and M.S. Joshi, 2013. Dual population Genetic Algorithm (GA) versus Open MP GA for multimodal function optimization. Int. J. Com. Applic., 64: 29-36.

Van Veldhuizen, D.A. and G.B. Lamont, 2000. On measuring multiobjective evolutionary algorithm performance. Proceedings of the Congress on Evolutionary Computation, Volume 1, July 16-19, 2000, La Jolla, CA., USA., pp: 204-211.

Wang, X. and O. Gotoh, 2010. A robust gene selection method for microarray-based cancer classification. Cancer Informat., 9: 15-30.

Witten, I. and E. Frank, 2005. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, USA.

Yu, L. and H. Liu, 2004. Efficient feature selection via analysis of relevance and redundancy. J. Mach. Learn. Res., 5: 1205-1224.