Volume 66, Issue 1, 2022
Journal of Scientific Research
of
The Banaras Hindu University

Feature Selection using PMOGA for Microarray Datasets

Seema Rathee[*1], Saroj Ratnoo[2], and Jyoti Ahuja[3]

[*1]Guru Jambheshwar University of Science & Technology, Hisar (India), Email ID (seema27rathee@gmail.com)
[2]Guru Jambheshwar University of Science & Technology, Hisar (India), Email ID (ratnoo.saroj@gmail.com)
[3]Government Post Graduate College for Women, Rohtak (India), Email ID (kwatra.jyoti@gmail.com)

Abstract: **Microarray technology is significantly impacting the community to know the primary characteristic underlining the expansion and growth of genes. Regardless of its many helpful relevance in analysis of drug detection and disease diagnosis; microarray data has turned into a dispute for various bio-analysts. The dimensionality problem in Microarrays leads to expansion of some new methods. The dilemma of dimensionality reduction in terms of features has been taken as a multi-objective optimization problem, thus, can be solved by using some multi-objective optimization techniques.**

Yet reduction in features takes a lot of time for final submission, therefore, parallel genetic algorithms will do this task in a more efficient way by parallel optimization of multiple distinct parts of a dataset. In this paper, we have designed a new combined method for parallel implementation of gene selection in multi-objective perspective named as PMOGA. Individual migration strategy is followed to improve the parallel searching speed for improving the efficiency of the proposed algorithm. A comparative study of the proposed PMOGA has been done on eight most referenced datasets. The obtained results confirm the supremacy of MOGA based parallel approach over the other approaches based on different performance measures.

Index Terms: **Feature selection (FS), KNN (K Nearest Neighbor) classifier, Microarray, Multi-objective genetic algorithm (MOGA) and Parallel Genetic Algorithm (PGA).**

I. INTRODUCTION

Microarrays are two dimensional arrays consist of samples (rows) and genes (columns). Microarrays are used to discover the point on which these genes are turned on or off in some cells and tissues. Distraction or any type of changes at several step of gene expression is prone for many inherent diseases. These thousands of genes can be calculated efficiently using microarrays. Therefore, the microarray technology has not only gives power to the society to know the life growth and development, but also has the potential to discover the genetic reasons of human body anomalies. It is proficient to find the causes of cancer by discovering the transformation in the sequences of genes. In view of the fact that genes verified the response of human bodies towards drugs, it can also used to recommend a drug treatment appropriate for a particular human being.

Generally, microarrays are exceptionally high dimensional with a very small sample size. Thus, microarrays consists of comprehensively large feature space which suffer from the nuisance of dimensionality problem (Yen, 2010). It becomes a challenge for data mining researchers to handle such a huge feature space. Despite of this dimensionality problem, microarray datasets contain noisy, irrelevant and redundant features which can badly influence the data mining algorithms. As a result, the selection of more significant and valuable features is of highest importance.

Besides all of the commonly used feature selection techniques like F-measure, T-test, mutual correlation-based feature selection, entropy-based feature selection etc, many methods have commenced diminishing the feature space by removing irrelevant and noisy features. Evolutionary Algorithms (Genetic Algorithms) and Swarm intelligence meta-heuristics have become very trendy in recent years. They have been widely used in the data mining community.

Literature confirms feature selection as a multi-objective problem. Several authors have applied multi-objective meta-heuristics for feature subset selection (Ahuja & Ratnoo, 2015; Anusha & Sathiaseelan, 2015; Grandchamp et al., 2015; Khan & Baig, 2015; Saroj & Jyoti, 2014; Spolaôr et al., 2017; Spolaor et al., 2010; Xue et al., 2014). The authors were successful in producing multiple feature subsets instead of a single best subset. Although authors have effectively attempted for feature selection using multi-objective optimization techniques, there is

a lack of research work based on parallel variants of multi-objective optimization algorithms (Natarajan, 2016; Natarajan & Balasubramanian, 2016).

Here, a parallel counterpart of the multi-objective genetic algorithm has been implemented to select many microarray data features. Non-Dominated Sorting Genetic Algorithm (NSGA II) commonly used MOGA to select features in a parallel setting. The individual migration and individual update strategy are initiated to keep the better convergence and diversity of the Pareto optimal set.

Rest of the work is structured as follows: Section 2 describes the related work for feature selection and multi-objective optimization. Section 3 gives detailed discussion of the proposed method. Experimental analysis and results are shown in section 4. Last, section 5 concludes the whole work.

II. Related Work

Feature selection is the principal necessity for analysis of microarray data. These dataset contains a number of unimportant and redundant features which need to be reduced. The task of feature selection is done by using some reduction techniques such as evolutionary or non-evolutionary algorithms. Evolutionary algorithms play a significant role for the purpose of feature selection (Ding & Liu, 2009; Dreyer, 2013; Goswami et al., 2018; Jović et al., 2015; Peralta et al., 2015; Tan et al., 2008; Xue et al., 2016). There are multiple contradictory objective functions for FS such as cost, time, and accuracy and reduction rate. Therefore, Feature selection is considered as multi-objective optimization problem. Multi-objective evolutionary algorithm optimizes multiple objective functions simultaneously (Ahuja & Ratnoo, 2015; Anusha & Sathiaseelan, 2015; Grandchamp et al., 2015; Khan & Baig, 2015; Saroj & Jyoti, 2014; Spolaor et al., 2010; Spolaôr et al., 2017; Xue et al., 2014).

The parallel versions of genetic algorithms take much less time as compared to simple genetic algorithm because they are executed on smaller parts of the dataset simultaneously. Some of the researchers have tried parallel genetic algorithm in different areas (Cano et al., 2011; Chen et al., 2016; Li & Huang, 2012; Silva et al., 2015). Chen et al. (2016) (Chen et al., 2016) has used a CGPGA (coarse-grained parallel genetic algorithm) to equally select feature subset and optimize parameters for SVM classifier in considerably a smaller amount of time. Adi and Aldasht (2018) (Adi & Aldasht, 2018) have categorized the algorithms in four groups i.e. Genetic Algorithms (GA), Scattered Search (SS), Particle Swarm Optimization (PSO) and Ant Colony Optimization (ACO). They have proved from their results that mainly PGAs (Parallel Genetic Algorithms) are suitable choice in terms of accuracy rather than parallel ACO for the problem of feature selection.

In a single objective optimization, the algorithms' execution time is very high due to sequential computation. The solution is parallel multi-objective genetic algorithm i.e. PMOGA. Some authors have implemented parallel and multi-objective algorithms simultaneously for their different purposes. A new hybrid method for feature selection has been designed, combining parallel and multi-objective techniques. In a multi-objective optimization approach, the best solutions come across every generation. They are transferred into the Pareto archive and then the solutions are selected from the Pareto archive during the construction procedure. This procedure is called elitism. According to the replacement strategy, the offsprings' solutions will replace their parents. This procedure repeats for a fixed number of time constraints given by users.

III. The Proposed Parallel Multi-Objective Genetic Algorithm for Feature Selection: PMOGA

In this paper, we have implemented Parallel MOGA for feature selection. Both parallel and multi-objective genetic algorithm used together for reducing various irrelevant features from the dataset. As we discussed earlier, NSGA II is a well-organized and perfect multi-objective genetic algorithm and can performs the genetic operations on a single population.

In this study, a coarse parallel island model has been designed from a multi-objective perspective. A different parallel multi-objective strategy has been used in this model, which combines these two mechanisms into a single framework to select features. This model is based on two populations' i.e. Elite population (E_Pop) and Searching population (S_Pop). To archive the non-dominated solutions in the whole population, an Elite population is used. It performs genetic operations on the Elite population separately as well as concurrently with the entire population. The second population is searching. The individual estimation and different genetic operations are carried out simultaneously in both populations.

This model best describes its migration strategy and individual update strategy. An individual migration strategy is shown in section 3.1. A detailed discussion of the proposed method is given in section 3.2. Section 3.3 specified the individual update strategy. Parameters and fitness functions are shown in section 3.4. Performance measures are explained in section 3.5. The last section 3.6, points out the selection method of the final best solution.

3.1. Individual Migration Strategy

A new mechanism i.e. migration strategy is introduced in the parallel multi-objective genetic algorithm. The migration is performed among the individuals of populations in the evolutionary process. The migration will increase the quantity of better individuals in each population to boost the convergence speed and improve classification accuracy. By using this migration strategy, the parallel algorithms can arrive at a single

population algorithm where parallel computing is driven by a single processor serially.

Both E_Pop and S_Pop develop at a similar time in PMOGA. At the evolution time of each generation, S_Pop sends its non-dominated individuals to E_Pop after applying both the mutation and crossover operators. Subsequently, E_Pop sends its dominated individuals to S_Pop at the same time. So the best individuals are then optimized in E_Pop. Alternatively, the individual's transfer from E_Pop to S_Pop will increase the speed of convergence. There is strict instruction for the individuals that similar individuals can migrate at most one time at individuals' migration time.

Figure 1 depicts the whole evolutionary process of PMOGA (left to right), including the migration policy. In Figure 1, p1 and p2 indicate E_Pop and S_Pop.

3.2. Detailed Discussion

A detailed discussion about the migration of individuals is given below:

Step1: Use p1 and p2 as populations for crossover, selection and mutation operations to form two new populations' np1 and np2.

Step2: Q1= p1Up2, Q2=np1Unp2. Pareto-domination is used to divide the population Q2 into two subparts; one contains the non-dominated individuals while another contains dominated. The plain box areas depict the non-dominated individuals in the figure, whereas the shaded area represents the dominated individuals.

Step3: In the non-dominated portion of Q2, transferring into a population Q1 is dissimilar from every chromosome in the population p1.

Step4: According to the Pareto-dominance idea, divide the population Q1 into two subparts as done in Step2, one consists of the non-dominated individuals and the other part contains dominated individuals. The differing individual of population p2 and Q2 are sent from population Q1 to population Q2.

Step5: Consider the non-dominated part of Q1 as the Elite Population p3 at the subsequent generation. The truncation/alteration strategy used in NSGA II is applied if the Elite population's size exceeds the maximum limit. Simultaneously, the truncation strategy is also applied to population p4 to generate Searching Population p4 for the next generation.

3.3. Individual Update Strategy

In this paper, a new strategy has been designed for updating the individuals. PMOGA uses searching Population to discover the best possible solution in the solution space. It sends all the non-dominated individuals originated from each generation to the Elite Population. If the number of chromosomes derived from the S_Pop to E_ Pop is found to be 0, then it is assumed that no non-dominated entity is found at this stage. If it is 0 up to several generations of GA, there is a need to update the S_Pop using an individual's update strategy.

3.4. Parameters and Fitness Function used by PMOGA

The parameters used for PMOGA for feature selection are population size, chromosome length, number of generations, crossover and mutation rate and tour size. The values set for the given parameters are almost same for both the populations i.e. E_pop and S_pop except crossover and mutation rate. The crossover rate for E_pop is 0.7 and 0.8 for S_pop. Population size=30, chromosome length according to number of features in the corresponding dataset, generations=40 and tour size=2.

The two most popular objectives i.e. classification accuracy and reduction rate of features are used in the fitness function where f1 uses as the accuracy in the fitness function and f2 as the reduction rate (Equation 1). The terms RS and DS are the taken as cardinality of the dataset.

$$f1 = \text{Accuracy and } f2 = \left(1 - \frac{|RS|}{|DS|}\right) * 100 \quad (1)$$

3.3. Performance Measures

For measuring the performance of various algorithms, we need some measure like sensitivity, specificity, G-Mean, PPV (Positive Predictive Value) and F-measure. All of these are given below in Table 1.

Table 1. Performance Measures

$\text{Sensitivity(Recall)} = \frac{TP}{TP + FN}$	(2)
$\text{Specificity} = \frac{TN}{TN + FP}$	(3)
$\text{GMean} = \sqrt{\text{Sensitivity} * \text{Specificity}}$	(4)
$\text{Precision} = \frac{TP}{TP + FP}$	(5)
$\text{F1 Score} = 2 * \left(\frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}\right)$	(6)

In the above table, TP refers to the true positive rate i.e. the positive tuples, which are precisely categorized by the classifier. TN is a true negative rate, which indicates the negative tuples that are accurately labeled by the classifier. FP is the false positive rate which signifies negative tuples that are wrongly labeled as positive. FN is the false-negative rate, which means

positive tuples that are inaccurately labeled as unfavorable by the KNN classifier [1].

3.6. Final Best Solution Selection

Multiple non-dominated solutions are arrived from the Pareto optimal set. The best solution is selected from the non-dominated solution set according to the users' requirements. There are several criteria for choosing a single best solution from the entire Pareto set. In this study, the best solution is selected from the instances common than the average number of solutions. The features that are regular in more than the average number of solutions include the final solution.

IV. Experimental analysis

4.1 Datasets

This section estimates the performance of the proposed PMOGA for gene selection from microarray datasets. Eight well-known real-life datasets have been used. Some of them are publically accessible medical domain datasets and are taken from the website http://www.biolab.si/supp/bi-cancer/projections/info. The experimentations were performed in a MATLAB environment. The detailed view of these datasets is given in Table 2.

Table 2. Summary of Datasets

Sr. No.	***Datasets***	***#Features***	***#Instances***	***#Classes***
1.	WDBC	30	569	2
2.	Lung cancer	19993	187	2
3.	Leukemia	7129	72	2
4.	Prostate Cancer	2135	102	2
5	DLBCL	7070	77	2
6.	GSE 412	8280	110	2
7	GSE 2535	12625	28	2
8.	GSE 2443	12627	20	2

4.2 Benchmark Techniques

Various traditional feature selection methods exist in the literature based on multi-objective optimization. DWFS: A wrapper feature selection tool based on a parallel genetic algorithm (Soufan et al., 2015) and Dimension reduction for microarray data using multi-objective ant colony optimization (Ahuja & Ratnoo, 2017) have been used as benchmark techniques to discover the performance of the novel proposed method. Soufan et al. (2015) have implemented the wrapper model and applied parallel GA which simultaneously evaluates massive features. DWFS incorporate diverse filtering methods and is used as a pre-processing step in feature selection (Soufan et al., 2015). Ahuja and Ratnoo et al. (2017) have designed a multi-objective ant colony optimization (MOACO) algorithm to select genes in which several non-dominated solutions are selected instead of a single solution (Ahuja & Ratnoo, 2017). Here, a broad comparison of the proposed algorithm PMOGA has been made with these two previous approaches i.e. DWFS and MOACOGS.

4.3 Results and Discussion

Since the wrapper methods usually take more time than the conventional methods. Thus, a time comparison of the proposed method with other traditional methods does not make a good deal. Hence, a suitable comparison has been made among PMOGA, DWFS and MOACOGS techniques.

4.3.1 Comparison between PMOGA and DWFS

The approach suggested in this paper is compared with DWFS and applied on only four datasets related to different medical datasets i.e. WDBC, Prostate, Leukemia and Lung cancer.

Soufan et al (2015) (Soufan et al., 2015) has compared DWFS with the commonly used filtering methods: 1) minimum redundancy maximum relevance (mRMR) (Hanchuan Peng et al., 2005), 2) joint mutual information (JMI) (Yang & Moody, 1999), 3) conditional mutual information maximization (CMIM) (Fleuret, 2004) and 4) interaction capping (ICAP) (Jakulin, 2005). DWFS select features using wrapper method of feature selection to estimate the efficiency of the pre-processing which is denoted as DWFS, wrapper FS joined with mRMR denotes mRMR+DWFS and wrapper FS united with JMI denotes JMI+DWFS. They have also compared the performance of DWFS and its deviations with the most efficient filtering and wrapper approaches, particularly, forward and backward search, sequential forward floating search (SFFS). The classification performance is calculated using five metrics of classification such as- sensitivity, specificity, GMean, Positive Predictive Value (PPV) and F1-measure (as shown in the equation given in section 3.5). A comparison has been made with DWFS and its variants. The results show that our approach is better than others. The performance is calculated using five metrics i.e. Sensitivity, Specificity, GMean, F1 measure and Positive Predictive Value (as shown in Tables 3 to 6).

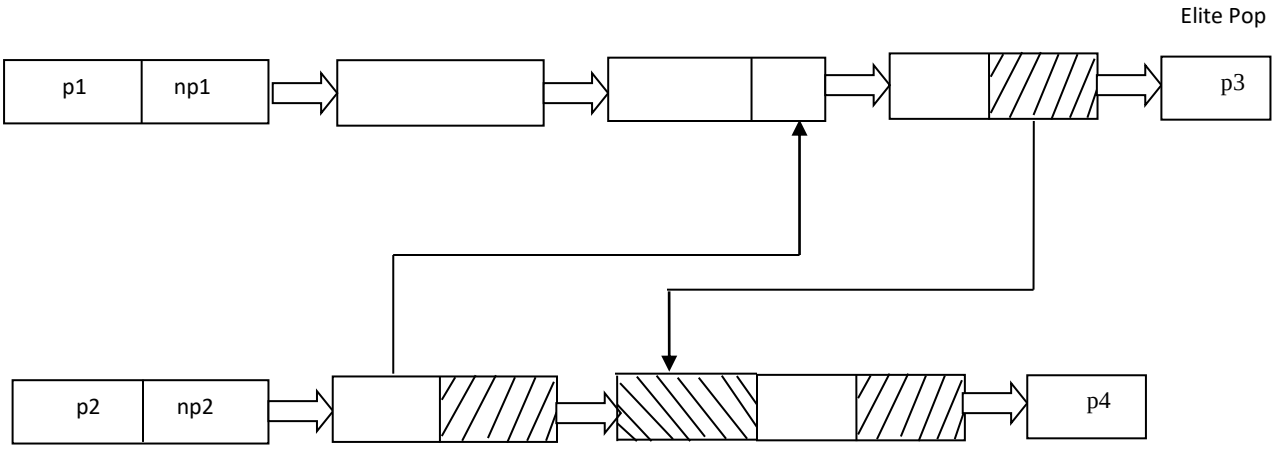

Fig.1. Migration Strategy of Individuals

Table 3. Classification performance metric using KNN classifier for WDBC dataset

	Sensitivity	**Specificity**	**G-Mean**	**PPV**	**F1-Measure**
PMOGA	94.23%	88.86%	91.50%	**98%**	**96.08%**
DWFS	96.86%	89.38%	93%	93.7%	95.2%
MRMR+DWFS	91.38%	85.61%	88.32%	91.35%	91.19%
JMI+DWFS	91.38%	85.61%	88.32%	91.35%	91.19%
MRMR	91.66%	85.61%	88.47%	91.37%	91.36%
JMI	95.31%	88.44%	91.8%	93.41%	94.33%
WEKA	92.2%	86.53%	89.29%	91.8%	91.93%
FST3	95.5%	87.92%	91.57%	92.9%	94.11%
ALL Features	96.16%	88.57%	92.27%	93.43%	94.75%
Correlation-Baseline	91.3%	84.05%	87.52%	90.19%	90.61%

Table 4. Classification performance metric for Lung cancer dataset using KNN classifier

	Sensitivity	**Specificity**	**G-Mean**	**PPV**	**F1-Measure**
PMOGA	**75%**	63.33%	**68.91%**	60%	**66.67%**
DWFS	63.52%	66.39%	64.82%	63.49%	63.14%
MRMR+DWFS	61.82%	66.72%	64.1%	62.88%	61.94%
JMI+DWFS	56.14%	78.33%	65.65%	70.06%	61.27%
MRMR	57.38%	71.94%	63.98%	66.19%	60.84%
JMI	54.96%	72.17%	62.43%	64.29%	58.41%
WEKA	61.66%	63.06%	62.04%	60.8%	60.26%
FST3	57.63%	67.06%	61.82%	61.37%	58.88%
ALL Features	59.43%	63.67%	60.89%	61.62%	59.65%
Correlation-Baseline	60.29%	65.44%	62.5%	61.44%	60.19%

Table 5. Classification performance metric using KNN classifier for Leukemia dataset

	Sensitivity	**Specificity**	**G-Mean**	**PPV**	**F1-Measure**
PMOGA	96%	85.91%	**90.81%**	**96%**	**96%**
DWFS	97.78%	85%	90.32%	91.92%	94.3%
MRMR+DWFS	94.29%	45.5%	61.74%	74.16%	81.57%
JMI+DWFS	90.81%	90.17%	90.18%	92.92%	91.37%
MRMR	92.78%	23%	35.98%	69.88%	78.87%
JMI	89.29%	68.67%	77.39%	81.72%	84.42%
WEKA	96.52%	86.17%	90.68%	91.51%	93.59%
FST3	95.96%	75.17%	84.47%	89.56%	92.57%
ALL Features	97.78%	82.67%	89.76%	89.11%	93%
Correlation-Baseline	98.33%	90%	93.79%	93.33%	95.4%

Table 6. Classification performance metric using KNN classifier for Prostate dataset

	Sensitivity	**Specificity**	**G-Mean**	**PPV**	**F1-Measure**
PMOGA	95.46%	89.66%	92.51%	88%	91.58%
DWFS	87.56%	85.33%	86.17%	84.57%	85.59%
MRMR+DWFS	94.42%	92.05%	93.17%	92.64%	93.4%
JMI+DWFS	98.18%	86.47%	92.03%	87.44%	92.3%
MRMR	88.93%	88.41%	87.99%	90.31%	88.54%
JMI	85.8%	84.65%	85.14%	84.39%	84.92%
WEKA	75.83%	83.51%	79.29%	79.85%	77.48%
FST3	84.48%	84.12%	84%	83.63%	83.88%
ALL Features	79.58%	83.11%	81.2%	81.02%	79.97%
Correlation-Baseline	92.25%	87.15%	89.44%	86.52%	88.84%

Table 7. Performance metrics for Prostate Dataset

	Sensitivity	**Specificity**	**Accuracy**	**F-score**	**Time (sec)**
PMOGA	0.9724	O.8966	0.9607	0.9362	94
MOACOGS	0.9733	0.9097	0.9268	0.9260	132
MOPSO	0.93459	0.912	0.9235	0.9265	351.09
SFS	0.89998	0.864	0.88234	0.88697	235.082
T Test	0.9269	0.816	0.87256	0.88132	112.49
Ranksum Test	0.91922	0.88	0.90002	0.9036	141.09
mRMR(MID)	0.9231	0.82	0.8725	0.8725	76.8
mRMR(MIQ)	0.8942	0.8916	0.9069	0.9069	64.8
CFS	0.9131	0.9201	0.9112	0.9211	233.9
CBFS	0.8558	0.93	0.8971	0.8971	61.599

Table 8. Performance metrics for DLBCL Dataset

	Sensitivity	**Specificity**	**Accuracy**	**F-score**	**Time (sec)**
PMOGA	0.94.37	0.9677	0.9489	0.9412	98
MOACOGS	0.9807	0.9312	0.9677	0.9807	126
MOPSO	0.92222	0.9379	0.9332	0.87635	237.42
SFS	0.74445	0.9655	0.9131	0.79647	28.8
T Test	0.83335	0.9172	0.89738	0.8035	103.31
Ranksum Test	0.8889	0.93449	0.9238	0.84979	127.181
mRMR(MID)	0.3035	0.9313	0.7961	0.3428	59.78
mRMR(MIQ)	0.3055	0.9483	0.8036	0.3428	80.11
CFS	0.5556	0.9355	0.8684	0.6667	51.482
CBFS	0.1944	0.9555	0.7829	0.2966	17.2112

Table 9. Performance metrics for GSE412 Dataset

	Sensitivity	**Specificity**	**Accuracy**	**F-score**	**Time (sec)**
PMOGA	0.8889	0.8929	0.9091	0.8889	132
MOACOGS	0.8928	0.9285	0.9090	0.8745	201
MOPSO	0.716	0.89667	0.81453	0.77904	445.39
SFS	0.68	0.9067	0.80363	0.75478	317.5
T Test	0.672	0.82668	0.7563	0.71157	386.34
Ranksum Test	0.7	0.89333	0.80544	0.76257	332.11
mRMR(MID)	0.5600	0.7825	0.7909	0.6968	121.73
mRMR(MIQ)	0.6200	0.7462	0.8136	0.7506	97.88
CFS	0.6400	0.9133	0.7990	0.7442	276.44
CBFS	0.7100	0.6359	0.7773	0.7427	46.67

Table 10. Performance metrics for GSE2535 Dataset

	Sensitivity	**Specificity**	**Accuracy**	**F-score**	**Time (sec)**
PMOGA	0.84	0.7500	0.6285	0.6333	76
MOACOGS	0.9166	0.5833	0. 79	0.7846	102
MOPSO	1	0.44447	0.80357	0.8585	269.53
SFS	0.84375	0.62499	0.74998	0.7897	272.32
T Test	0.71875	0.625	0.6786	0.69905	101.269
Ranksum Test	0.75	0.5834	0.67857	0.70533	117.153
mRMR(MID)	0.3750	0.8331	0.5714	0.5000	34.186
mRMR(MIQ)	0.6250	0.7722	0.7143	0.7143	36.245
CFS	0.5900	0.8771	0.7143	0.6967	232.06
CBFS	0.6250	0.7343	0.7143	0.7143	39.16

Table 11. Performance metrics for GSE2443 Dataset

	Sensitivity	**Specificity**	**Accuracy**	**F-score**	**Time (sec)**
PMOGA	0.8667	0.8500	0.90	0.9273	85
MOACOGS	1	0.99	1.0	1	132
MOPSO	1	0.96	0.98	0.981818	287.92
SFS	0.84	0.92	0.88	0.8723	239.482
T Test	0.92	0.88	0.9	0.9094	121.47
Ranksum Test	1	0.96	0.98	0.98182	176.778
mRMR(MID)	1	0.8	0.9	0.9091	87.43
mRMR(MIQ)	0.987	0.82	0.9001	0.9091	83.611
CFS	1	0.8010	0.9021	0.9091	205.25

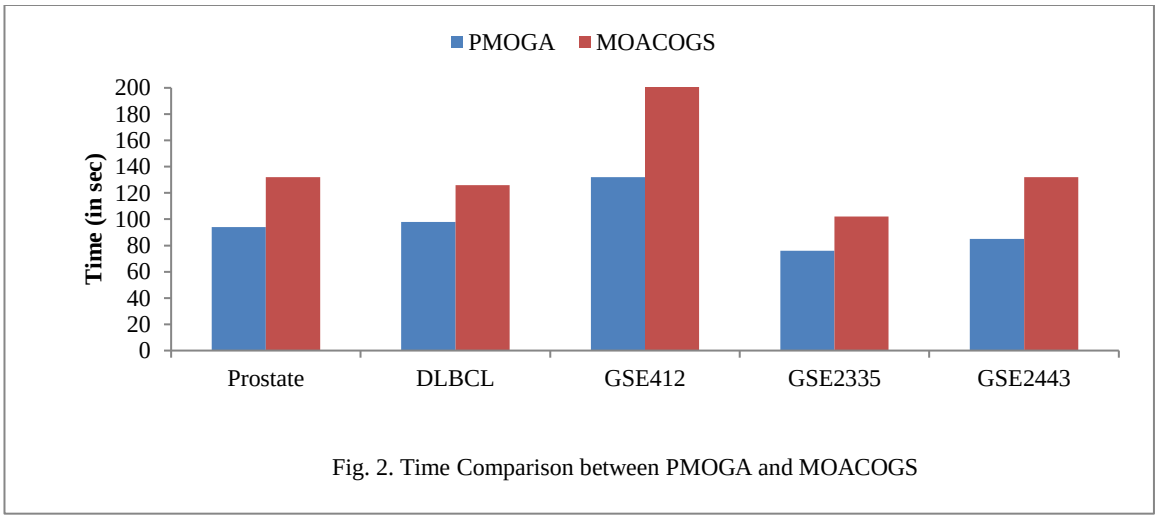

Fig. 2. Time Comparison between PMOGA and MOACOGS

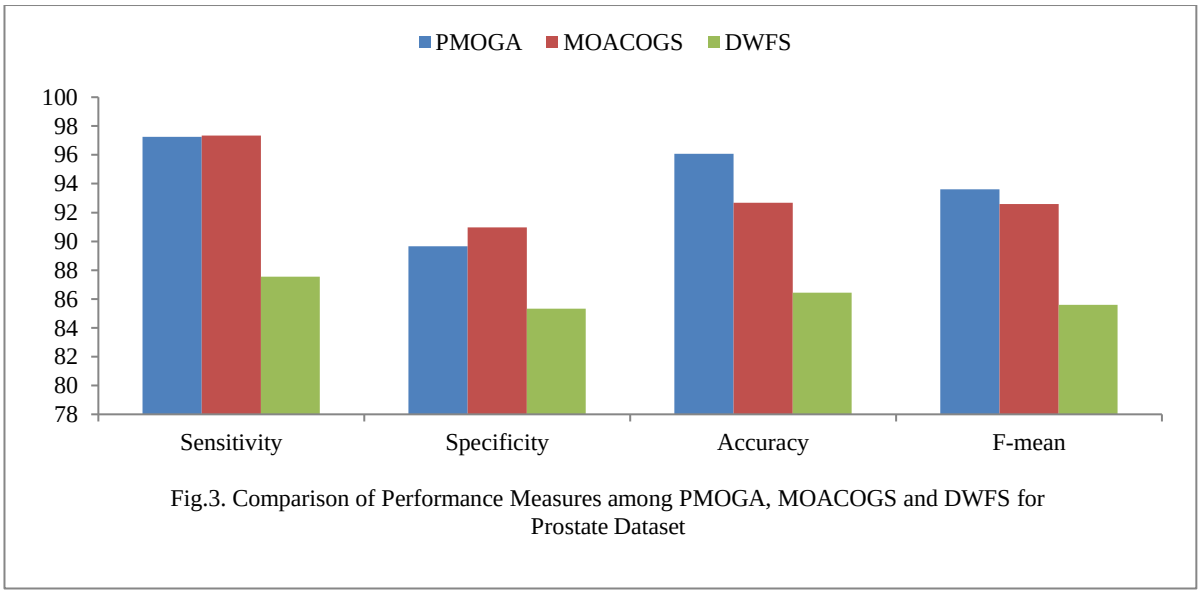

Fig.3. Comparison of Performance Measures among PMOGA, MOACOGS and DWFS for Prostate Dataset

In lung cancer datasets, the sensitivity measured by our approach is much better than other methods. The specificity, G-mean measures describes that the proposed approach PMOGA is comparable to all the other approaches for all the datasets. For Leukemia and WDBC datasets, the PMOGA is far better than the rest of the measuring positive predictive value. In the F1 measure, our proposed method is far better than earlier approaches for WDBC, Leukemia and Lung Cancer datasets except for the leukemia dataset.

Thus, it is clear from the compared results shown in Tables 3 to 6 that PMOGA outperforms DWFS and its subsequent approaches in terms of all the performance metrics used.

4.3.2 Comparison between PMOGA and MOACOGS

The performance of PMOGA is tested on five datasets. The comparison is based on the following datasets- Prostate, DLBCL, GSE 412, GSE 2535 and GSE 2443.

The proposed algorithms' results are calculated according to the result format of the compared algorithm (MOACOGS).

The performance is measured using these metrics:-sensitivity, specificity, accuracy, f-score and execution time. Tables 7 to 11 illustrate the comparison of the proposed approach and MOACOGS with some benchmark techniques. When the proposed method is compared with MOACOGS, it is clearly visible that our approach is comparative to MOACOGS in almost all the datasets. To some extent, MOACOGS is better, but PMOGA is better in terms of execution time for sure.

There is clear supremacy of the proposed method over the traditional methods for almost all the datasets regarding classification accuracy. If the comparison is made based on average sensitivity and specificity, the proposed algorithm is comparable to all the other methods. However, the difference is not significant. Similarly, all the other methods dominate over the MOACOGS in GSE 2535 (Chronic Myeloid Leukemia Treatment) except MOPSO based method. A comparison made on f-scores also reveals that our algorithm is analogous to the other methods.

As we know, the wrapper techniques generally take more time than the traditional methods. Thus, a time comparison of the proposed method with other classical methods does not make any sense. Hence, the only apt comparison among our proposed method, MOACOGS and MOPSO methods, is based on time. MOACOGS is a multi-objective approach that takes additional time for execution. For removing the limitation of this approach, a parallel multi-objective approach (PMOGA) for data reduction is designed.

As mentioned earlier, applying multi-objective parallel schemes would produce an incredible impact over former core evolutionary approaches. To validate this claim, Tables 7 to 11 show each method's execution time for each of the datasets. The time is calculated from the starting of the algorithm awaiting its final output. The result indicates from Figure 2 that PMOGA is significantly faster than MOACOGS.

In Figure 3, a graphical comparison has been made among the three approaches PMOGA, MOACOFS and DWFS, in terms of sensitivity, specificity, accuracy and f-mean for the Prostate dataset. The results show the supremacy of the proposed approach i.e. PMOGA, for most of the performance measures.

V. Conclusion

This paper presents the feature selection problem from large dimensional microarray datasets as a multi-objective optimization problem. A multi-objective PGA-based method for feature selection has been proposed. Precisely, a combined MOGA and PGA strategy has been followed to design the algorithm (PMOGA) to extract the most valuable features. Eight real-life microarray technology-based datasets have been chosen for confirmation of the proposed approach.

The performance of our proposed approach is compared with some traditional feature selection methods. The results confirm that the proposed algorithm is either better or comparable to other algorithms across all the datasets on several performance metrics except minor exceptions. From the experimentation, it has been concluded that the proposed multi-objective genetic algorithm for feature selection runs significantly faster than both the algorithms i.e. MOACOGS and MOPSO. The hybridization of both MOGA and PGA gives extensive feature selection results and discovers multiple Pareto optimal solutions instead of a single solution. Therefore, a user can choose the best optimal solution according to his/her preferences.

In future, we can consider parallel MOGA for the problem of dual (instances as well as features) selection problem for other disease related microarray datasets.

References

Adi, S. I., & Aldasht, M. (2018). A. *American Journal of Computer Science and Engineering Survey*, *6*(1), 013–021.

Ahuja, J., & Ratnoo, S. (2017). Dimension reduction for microarray data using multi-objective ant colony optimisation. *International Journal of Computational Systems Engineering*, *3*(1–2), 58–73. https://doi.org/10.1504/IJCSYSE.2017.083149

Ahuja, J., & Ratnoo, S. D. (2015). Feature Selection using Multi-objective Genetic Algorithm: A Hybrid Approach. *ACM Recommender System 2012 Workshop on*

Recommendation Utility Evaluation: Beyond RMSE, Dublin, Ireland, *14*(1), 26–37.

Anusha, M., & Sathiaseelan, J. G. R. (2015). Feature Selection Using K-Means Genetic Algorithm for Multi-objective Optimization. *Procedia Computer Science, 3rd International Conference on Recent Trends in Computing 2015 (ICRTC-2015)*,*57*,1074–1080. https://doi.org/10.1016/j.procs.2015.07.387

Cano, A., Zafra, A., & Ventura, S. (2011). A Parallel Genetic Programming Algorithm for Classification. In E. Corchado, M. Kurzyński, & M. Woźniak (Eds.), *Hybrid Artificial Intelligent Systems* (pp. 172–181). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-21219-2_23

Chen, Z., Lin, T., Tang, N., & Xia, X. (2016). A Parallel Genetic Algorithm Based Feature Selection and Parameter Optimization for Support Vector Machine. *Science Program*, *2016*, 1–11. https://doi.org/10.1155/2016/2739621

Ding, S., & Liu, X. (2009). Evolutionary Computing Optimization for Parameter Determination and Feature Selection of Support Vector Machines. *International Conference on Computational Intelligence and Software Engineering, 2009. CiSE 2009*, 1–5. https://doi.org/10.1109/CISE.2009.5366095

Dreyer, S. (2013). *Evolutionary Feature Selection* [Norwegian university of science and technology]. http://www.diva-portal.org/smash/record.jsf?pid=diva2%3A702890&dswid=1670

Fleuret, F. (2004). Fast Binary Feature Selection with Conditional Mutual Information. *Journal of Machine Learning Research*, *5*, 1531–1555.

Goswami, S., Chakrabarti, A., & Chakraborty, B. (2018). An empirical study of feature selection for classification using genetic algorithm. *International Journal of Advanced Intelligence Paradigms*, *10*(3), 305–326. https://doi.org/10.1504/IJAIP.2018.090792

Grandchamp, E., Abadi, M., & Alata, O. (2015). An Hybrid Method for Feature Selection Based on Multiobjective Optimization and Mutual Information. *Journal of Informatics and Mathematical Sciences*, *7*(1), 21–48. https://doi.org/10.26713/jims.v7i1.268

Hanchuan Peng, Fuhui Long, & Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(8), 1226–1238. https://doi.org/10.1109/TPAMI.2005.159

Jakulin, A. (2005). *Machine Learning Based on Attribute Interactions* [Phd, Univerza v Ljubljani]. http://eprints.fri.uni-lj.si/205/

Jović, A., Brkić, K., & Bogunović, N. (2015). A review of feature selection methods with applications. *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 1200–1205. https://doi.org/10.1109/MIPRO.2015.7160458

Khan, A., & Baig, A. R. (2015). Multi-Objective Feature Subset Selection using Non-dominated Sorting Genetic Algorithm. *Journal of Applied Research and Technology*, *13*(1), 145–159. https://doi.org/10.1016/S1665-6423(15)30013-4

Li, W., & Huang, Y. (2012). A Distributed Parallel Genetic Algorithm oriented adaptive migration strategy. *2012 8th International Conference on Natural Computation*, 592–595. https://doi.org/10.1109/ICNC.2012.6234584

Natarajan, A. (2016). *A Fuzzy Parallel Island Model Multi Objective Genetic Algorithm Gene Feature Selection For Microarray Classification.*

Natarajan, A., & Balasubramanian, D. R. (2016). A Parallel Multi Objective Optimization Genetic Algorithm Gene Feature Selection on Microarray Based Cancer Classification Using Neuro-Fuzzy Inference System. *International Journal of Scientific & Engineering Research*, *7*(3), 442–450.

Peralta, D., del Río, S., Ramírez-Gallego, S., Triguero, I., Benitez, J. M., & Herrera, F. (2015). Evolutionary Feature Selection for Big Data Classification: A MapReduce Approach. *Mathematical Problems in Engineering*, 1–11. https://doi.org/10.1155/2015/246139

Saroj & Jyoti. (2014). Multi-objective genetic algorithm approach to feature subset optimization. *2014 IEEE International Advance Computing Conference (IACC)*, 544–548. https://doi.org/10.1109/IAdCC.2014.6779383

Silva, J., Aguiar, A., & Silva, F. (2015). A Parallel Computing Hybrid Approach for Feature Selection. *2015 IEEE 18th International Conference on Computational Science and Engineering*, 97–104. https://doi.org/10.1109/CSE.2015.34

Soufan, O., Kleftogiannis, D., Kalnis, P., & Bajic, V. B. (2015). DWFS: A Wrapper Feature Selection Tool Based on a Parallel Genetic Algorithm. *PLOS ONE*, *10*(2), e0117988. https://doi.org/10.1371/journal.pone.0117988

Spolaôr, N., Lorena, A. C., & Lee, H. D. (2017). Feature Selection via Pareto Multi-objective Genetic Algorithms. *Applied Artificial Intelligence*, *31*(9–10), 764–791. https://doi.org/10.1080/08839514.2018.1444334

Spolaor, N., Lorena, A. C., & Lee, H. D. (2010). Use of Multiobjective Genetic Algorithms in Feature Selection. *2010 Eleventh Brazilian Symposium on Neural Networks*, 146–151. https://doi.org/10.1109/SBRN.2010.33

Tan, F., Fu, X., Zhang, Y., & Bourgeois, A. G. (2008). A genetic algorithm-based method for feature subset selection. *Soft Computing*, *12*(2), 111–120. https://doi.org/10.1007/s00500-007-0193-8

Xue, B., Fu, W., & Zhang, M. (2014). Multi-objective Feature Selection in Classification: A Differential Evolution Approach. In G. Dick, W. N. Browne, P. Whigham, M.

Zhang, L. T. Bui, H. Ishibuchi, Y. Jin, X. Li, Y. Shi, P. Singh, K. C. Tan, & K. Tang (Eds.), *Simulated Evolution and Learning* (pp. 516–528). Springer International Publishing.

Xue, B., Zhang, M., Browne, W. N., & Yao, X. (2016). A Survey on Evolutionary Computation Approaches to Feature Selection. *IEEE Transactions on Evolutionary Computation*, *20*(4), 606–626. https://doi.org/10.1109/TEVC.2015.2504420

Yang, H. H., & Moody, J. (1999). Data Visualization and Feature Selection: New Algorithms for Nongaussian Data. *Proceedings of the 12th International Conference on Neural Information Processing Systems*, 687–693.

Yen, N. (Ed.). (2010). *Advances in Computational Biology*. Springer-Verlag. https://www.springer.com/gp/book/9781441959126

***