Feature Selection Using Multi-Objective Optimization Technique for Supervised Cancer Classification

P. Agarwalla and S. Mukhopadhyay

1 Introduction

The importance of classifying cancer and appropriate diagnosis of advancement of disease has leveraged many research fields, from biomedical to the machine learning (ML) domains. For proper diagnosis of a disease and categorizing it into different classes, investigation in the changes of genetic expression level is needed. Gene expression data (Zhang et al. 2008) has a huge impact on the study of cancer classification and identification. The ability of machine learning approaches to detect key features from a huge complex dataset reveals their importance in the field of feature selection from microarray dataset. Modelling of cancer progression and classification of disease can be studied by employing learning-based approaches. The methodology that has been intimated here is based on supervised learning technique for different input feature genes and data samples. In the supervised learning process, a set of training data has been provided along with their class information. Based on the methodology, it will identify the relevant informative features which will further identify the class of an unknown test sample. Multi-objective optimization techniques are involved in this paper to select the required features which are efficient in the classification purpose as well as carry significant biological information related to disease. Then, those features are used to train the classifier and a new sample is diagnosed.

Different approaches are developed by the researchers for finding marker genes (Khunlertgit and Yoon 2013; Bandyopadhyay et al. 2014; Mukhopadhyay and Mandal 2014; Apolloni et al. 2016) related to different diseases. Various statistical filter

P. Agarwalla (✉)
Heritage Institute of Technology, Kolkata, India
e-mail: prativa.agarwalla87@gmail.com

S. Mukhopadhyay
Institute of Radio Physics & Electronics, Kolkata, India
e-mail: sumitra.mu@gmail.com

J. K. Mandal et al. (eds.), *Multi-Objective Optimization*,
https://doi.org/10.1007/978-981-13-1471-1_9

approaches (Bandyopadhyay et al. 2014), clustering processes (Mukhopadhyay and Mandal 2014) and wrapper-based hybrid approaches (Apolloni et al. 2016) are utilized for this purpose. Many supervised and unsupervised classification techniques are used for classification or clustering of tissue samples. A family of bio-inspired algorithms has also been applied while formulating the problem as an optimization problem. The gene subset identification problem can be reduced to an optimization problem consisting of a number of objectives. However, identifying most relevant and non-redundant genes is the main goal that is to be achieved. Motivated by this, different multi-objective methodologies are proposed in the literature (Mukhopadhyay and Mandal 2014; Ushakov et al. 2016; Zheng et al. 2016; Mohamad et al. 2009; Hero and Fluery 2002; Chen et al. 2014). Recently, a number of literature involve multi-objective based methods for the feature selection from microarray datasets. To obtain a small subset of non-redundant disease-related genes by using the multi-objective criterions, different bio-inspired algorithms are applied. For example, a variable-length particle swarm optimization (Mukhopadhyay and Mandal 2014) is implemented. A bi-objective concept is implemented for clustering of cancer gene from microarray datasets (Ushakov et al. 2016). In work (Zheng et al. 2016), a numerical method is implemented with GA to extract informative features in the domain of bioinformatics. Multi-objective function is optimized by genetic algorithm (GA) (Mohamad et al. 2009) to obtain significant genes for cancer progression. Pareto-based analysis is performed for filtering the relevant genes (Hero and Fluery 2002). In this chapter, the problem is formulated as a multi-objective optimization problem, and multi-objective blended particle swarm optimization (MOBPSO), multi-objective blended differential evolution (MOBDE), multi-objective blended artificial bee colony (MOBABC) and multi-objective blended genetic algorithm (MOBGA) are proposed for this purpose. Here, the stochastic algorithms are modified using Laplacian blended operator to incorporate diversity in the search process. It helps to get more diversified and promising result. This has been established theoretically and experimentally in the subsequent sections. The modified multi-objective algorithms are searching for Pareto-front solution which represents the feature genes for cancer classification. Then, the comparative analysis is performed based on the efficiency of finding relevant marker genes which are significantly associated with the disease.

For the reliable classification of a disease, multiple objectives play an important role. In the context of gene selection from the microarray data, two objectives are considered. One of them allows selection of the most differentially expressed genes which help in identifying the separation between classes. Another consideration is given to the accurate classification of the disease. T-score is used for the job of selecting differentially expressed feature. Those selected genes may not be efficient to provide good classification result due to the heterogeneous nature of gene expression. Our mission is to choose the combination of feature which is providing high accuracy also. Again, if entire differential features are used, it can cause over-fitting of the classifier. So, it is necessary to eliminate redundant features for the task of classification. Here, our aim is to obtain high accuracy of classification. As well as the selected features should have good value of t-score and this in turn indicates differentiability in expression level. If the proper combination of genes for the deter-

mination of disease can be identified, then it will be a significant contribution for the diagnosis of disease and the treatment will be more effective and precise.

Experiments are performed using different types of microarray datasets which include Child_ALL (Cheok et al. 2003), gastric (Hippo et al. 2002), colon (Alon et al. 1999) and leukemia (Golub et al. 1999) cancer data. Initially, the performance of the proposed methodologies for the job of classification of disease through supervised learning process is evaluated. As mentioned, in this chapter, differential evolution (DE) (Price et al. 2006), artificial bee colony (ABC) (Karaboga and Basturk 2007), genetic algorithm (Yang and Honavar 1998) and particle swarm optimization (PSO) (Kennedy 2011) algorithms are involved for solving multi-objective feature selection problem along with Laplacian operator. Then, the results of classification are compared with other established methods for four real-life cancer datasets. The proposed Laplacian operator integrated with multi-objective swarm and evolutionary algorithms establishes good results in all respect. The results ascertain the ability of multi-objective blended particle swarm optimization (MOBPSO), multi-objective blended differential evolution (MOBDE), multi-objective blended artificial bee colony (MOBABC) and multi-objective blended genetic algorithm (MOBGA) to produce more robust gene selection activity. At the end of the chapter, the biological relevance of the resultant genes is also validated and demonstrated.

The remaining of the chapter is presented as follows: First, a description of experimental datasets is presented. Next, the proposed technique is presented for marker gene selection. In the next section, the result of the proposed technique is demonstrated and a comparative analysis is provided. Finally, the biological relevance of the result is given.

2 Experimental Datasets

Two classes of raw microarray data for different types of cancers are collected. In microarray data, the expressions of genes are arranged column-wise, whereas the samples, collected from different sources, are arranged in row. The changes at molecular level of genes can be visualized from the microarray technology. Here, gene expressions from different samples are analysed in a single microscopic slide. Samples from cancerous and non-cancerous tissues are taken and dyed using fluorescent colours. Then, through hybridization procedure, the combined colours are analysed. The intensity of different areas of microarray slide reveals the informative content and subsequently, conclusion can be made by investigating the expression level. Authors have collected microarray datasets of different variants of cancers from reliable sources such as **National Centre for Biotechnology Information** (**NCBI**). A brief description of the microarray datasets used for the experimental purpose is given below.

Child-ALL (GSE412) (Cheok et al. 2003): 110 samples of childhood acute lymphoblastic leukemia are collected. Among them, 50 and 60 examples are of before

and after therapy, respectively. The samples are having expression level of 8280 genes. So, the dimension of the dataset matrix is $D_{110X8280}$.

Gastric Cancer (GSE2685) (Hippo et al. 2002): Gastric cancer occurs due to the growth of cancerous cells in the lining of stomach. This experimental dataset is having expression of 4522 genes from total 30 number of different tissue samples. Two classes of tumour such as diffuse and intestinal advanced gastric tumour samples are considered. 22 samples are present in the first class and another class is having 8 samples.

Colon cancer (Alon et al. 1999): The colon dataset contains expression values of 6,000 genes column-wise. Totally, 62 cell samples are present row-wise, among which first 40 biopsies are from tumour cells and next 22 samples are from healthy parts of the colon. The data is collected from a public available website: www.biconductor.com/datadet.

Lymphoma and Leukemia (GSE1577) (Golub et al. 1999): The leukemia dataset consists of 72 microarray experiments including two types of leukemia, namely AML (25 samples) and ALL (47 samples). Expressions of 5147 genes are present in the dataset. The data is collected from a public website: www.biolab.si/supp/bi-cancer/projections/info/.

Preprocessing Microarray Data: The microarray data generally consists of noisy and irrelevant genes which may mislead the computation. So, to extract most informative and significant gene subset which is relevant for the diagnosis of the disease, first the noisy and irrelevant genes are to be eliminated. To analyse the noise content, signal-to-noise ratio is calculated for each gene and based on the SNR value, the top 1000 genes are selected for the next level of computation. The formula for SNR value calculation is given in Eq. (1). Here, μ_1, μ_2 are the means of gene expression of a particular gene over the samples of first class and second class, respectively. sd_1, sd_2 are the standard deviations of gene expression of a particular gene over the samples of first and second class, respectively.

$$SNR = \frac{\mu_1 - \mu_2}{sd_1 + sd_2} \tag{1}$$

Next, using min-max normalization process (Bandyopadhyay et al. 2014), those 1000 genes are normalized. If the expression of a gene over the samples is represented by the variable g, then the min-max normalization formula for a data point g_i is described by Eq. (2). Thus, a data matrix D_{mx1000} is formed where m is the number of samples. This generated data matrix is used for the next level of computation.

$$x_i(normalized) = \frac{g_i - \min(g)}{\max(g) - \min(g)} \tag{2}$$

3 Objectives

A multi-objective optimization problem (MOP) can be represented as follows:

$$\text{maximize } F(x) = (f(x), \ldots, fm(x))^T \tag{3}$$

subject to $x \in \Omega$, where Ω is the search space and x is the decision variable vector. F: $\Omega \rightarrow \mathrm{R}^m$, where m is the number of objective functions, and R^m is the objective space. The heterogeneity in the expression level of genes must be high from one patient to another and an optimal combination of feature set through learning process is to choose which will perform well for the classification of new sample. It has been noticed that a particular combination of gene set which is highly differentiable from one class to another sometimes fails to achieve good classification result. It sometimes causes over-fitting of the classifier. So, to keep a balance between them, a multi-objective problem is constructed. It gives rise to a set of trade-off between Pareto-optimal (P-O) solutions (Srinivas and Deb 1994). Here, two objectives are considered in this chapter which are described below:

$$t - score = \frac{\mu_1 - \mu_2}{\sqrt{\left(\frac{sd_1^2}{n_1} + \frac{sd_2^2}{n_2}\right)}} \tag{4}$$

$$Accuracy = \frac{t_n + t_p}{t_n + t_p + f_p + f_n} \tag{5}$$

For t-score calculation, the mean expression of the selected genes over the samples for both the classes is calculated. Then, the difference between the two mean expressions is computed. For fitness function for PSO computation t-score is utilized which is described in Eq. (4) where μ and sd represent the mean and the standard deviation value of the two classes, respectively. n_1 and n_2 are the number of samples present in the two classes, respectively. Higher fitness function indicates the better selectivity of genes. As another objective function, accuracy is estimated using the number of false positive (fp), true negative (tn), false negative (fn) and true positive (tp) for class prediction. The objective used for formulating multi-objective problem and the proposed methodology is discussed in brief in the following sections.

4 Proposed Methodology

The problem has been modelled as a multi-objective optimization problem, and different multi-objective evolutionary algorithms are employed. In the multi-objective optimization problem, a set of solutions called Pareto-optimal has to be achieved. Here, based on two objective functions, optimal Pareto solution is generated and for this purpose non-dominated sorting technique (Srinivas and Deb 1994) has been

used. A new version of optimization algorithm is proposed and developed, entitled as multi-objective blended particle swarm optimization (MOBPSO) algorithm for finding gene subsets in cancer progression. PSO algorithm is modified, and Laplacian blended operator is integrated to provide better diversity in searching procedure. So, the multi-objective blended PSO based concept is implemented along with GA, DE and ABC algorithms and MOBABC, MOBGA and MOBDE are developed. Subsequently, a comparative study is performed using the proposed multi-objective stochastic computational methods. For the selection of genes from microarray data, supervised learning method is employed where the total experimental dataset is partitioned into two subsets. One is used for training purpose of the proposed model and the other one is used for evaluation of the model. The schematic diagram of the proposed methodology is shown for MOBPSO in Fig. 1, and the process selection of bio-markers from gene expression profile is described below.

4.1 Multi-Objective Blended Particle Swarm Optimization (MOBPSO)

4.1.1 Concept of Particle Swarm Optimization (PSO)

Particle swarm optimization, proposed by Eberhart and Kennedy in 1995, is a simple, well-established and widely used bio-inspired algorithm in the field of optimization. The technique is developed based on the social behaviour of a bird flock, as the flock searches for food location in a multidimensional search space. Location of a particle represents the possible solutions for the optimization function, f(x). Velocity and the direction of a particle are affected by its own past experience as well as other particles in the swarm have an effect on the performance. The velocity and position update rule for *i*th particle at *t*th generation are given in Eqs. (6) and (7) where the values of two random weights, c_1 and c_2, represent the attraction of a particle towards its own success p_{best} and the attraction of a particle towards the swarm's best position g_{best} respectively. *w* is the inertia weight. After a predetermined number of iterations, the best solution of the swarm is the solution of the problem.

$$v^i(t) = w * v^i(t-1) + c_1 * rand * \left(p^i_{best} - x^i\right) + c_2 * rand * \left(g^i_{best} - x^i\right) \quad (6)$$

$$x^i(t) = x^i(t-1) + v^i(t) \quad (7)$$

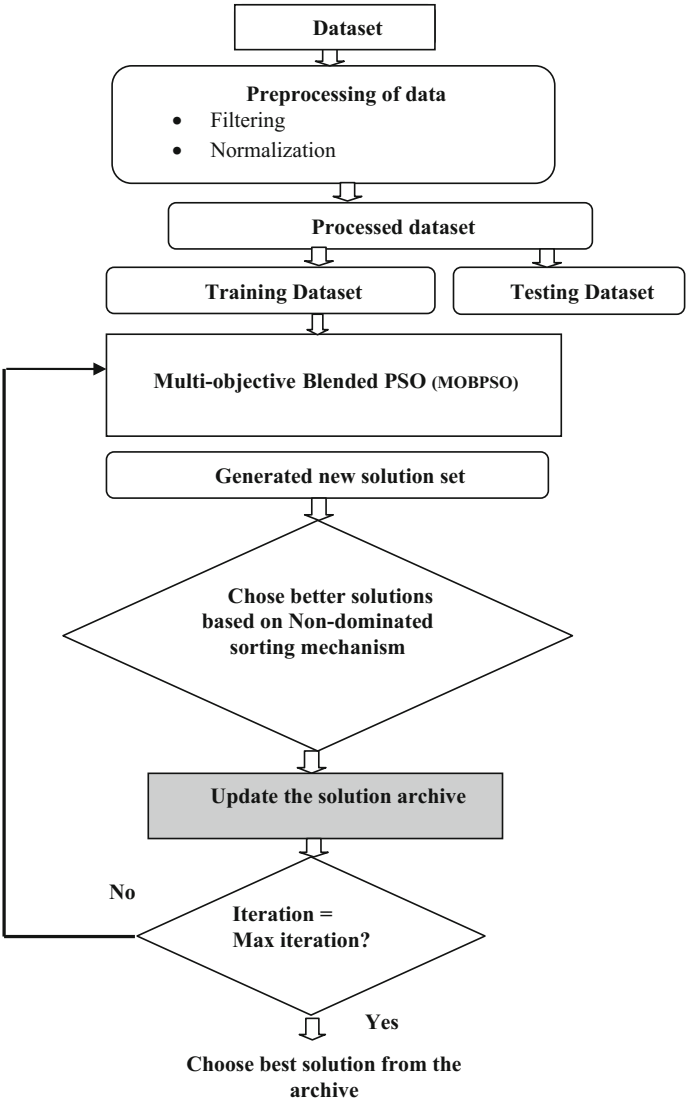

Fig. 1 Computational methods using MOBPSO-based approach

4.1.2 Concept of Multi-Objective Blended Particle Swarm Optimization (MOBPSO) for the Selection of Genes

The main drawback of PSO algorithm is that it is easily trapped to local optima due to scarcity of divergence which leads to premature convergence. To get rid of the issue, a diversity mechanism should be applied to get rid of any local optima. So,

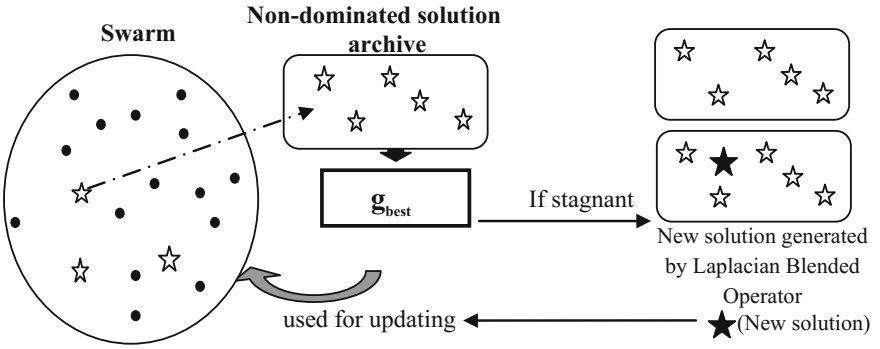

Fig. 2 Schematic diagram of the proposed methodology

in course of searching, better results can be achieved by introducing some sort of diversity technique. Being motivated by this, a blended operator is implemented with PSO and MOBPSO is proposed. In MOBPSO, particles are searching for optima, following the rules of PSO and if any particle is stuck at local optima, then it comes out of the situation by using the new probable solution generated through Laplacian blended operator. Blended Laplacian operator works very effectively to generate new probable solutions in the search space. The whole swarm is directed to the new solution which helps to discover new area of searching. As a whole the performance of the algorithm is accelerated and a better optimized result of the problem can be obtained. The mechanism is discussed below and shown in Fig. 2. MOBPSO is applied in the domain of multi-objective problem where the aim is to choose the Pareto-optimal solution. For the selection of feature genes, it is optimizing two objective functions and after each iteration non-dominated solutions are selected. As a single fitness value cannot be assigned, a modification is performed in the updating rule of the particles. In MOBPSO, the particles of the swarm are updating their velocity, and position towards the food using Eqs. (8), (9) and during updating the effect of g_{best} is only taken into consideration.

$$v^i(t) = w \cdot v^i(t-1) + c_1 \cdot rand \cdot (g^i_{best} - x^i) \tag{8}$$

$$x^i(t) = x^i(t-1) + v^i(t) \tag{9}$$

The g_{best} is the best solution chosen among the non-dominated solutions obtained so far. For the problem of identifying significant genes, the differentially expressed genes in different classes are important to be identified. As mentioned, t-score is used as one of the objective functions for the purpose. Accuracy is chosen as another objective function where the aim is to maximize the value of the accuracy. Now for each iteration, new subset of genes is generated by MOBPSO. The position of each particle represents a possible gene subset of the problem. Then, the fitness value of

each particle is calculated and based on the non-dominated sorting, better solutions are sorted. The non-dominated solutions are stored in an archive. As, for multi-objective problem, a number of Pareto-optimal solutions can be achieved, one of the Pareto-optimal solutions is randomly chosen as g_{best}. The g_{best} is used for the updating of velocity and position of a particle of MOBPSO. In the next iteration, again new subsets of genes are selected by the particles and the non-dominated solutions are loaded into the archive. Now, the archive is updated with the non-dominated solutions among the solutions obtained so far. A binary version of the optimization algorithm is used to select feature genes which are to be presented in the computation. The selection of genes is done based on Algorithm 1.

Algorithm-1 (Implementation of Binary concept)

```
for j=1:dimension of particle
  if x (j) > 0.5
     x(j)=1;
  else x (j)=0;
  end;
end;
```

During the search process, it may happen that generation of new better solution is stuck after few iteration due to the lack of diversity. So, the algorithm needs some mechanism which can direct the particles to a new probable region. Blended Laplacian operator works very efficiently to provide diversity to the swarm. If no better solution is generated, blended operator produces a new g_{best} at that point to provide diversification to the swarm. The mechanism is shown schematically in Fig. 2. First, two random solutions, sol_1 and sol_2, are chosen from the archive. Then, using a random coefficient termed as beta, two new solutions y_1 and y_2 are produced. The new g_{best}, g_{best_new}, is a combination of these two new solutions y_1 and y_2 having a weightage factor gamma. Blended Laplacian operator used for g_{best_new} generation is described below. g_{best_new} is completely a new solution generated from the old best non-dominated solutions, achieved so far. The new solution works to direct the entire swarm to a new possible direction.

$$\begin{aligned} gamma &= 0.1 + (1-0.1)^{0.95*iter} \\ beta &= 0.5 * log(rand) \\ y_1 &= sol_1 + beta * (sol_1 - sol_2) \\ y_2 &= sol_2 + beta * (sol_1 - sol_2) \\ g_{best\_new} &= gamma * y_1 + (1 - gamma) * y_2 \end{aligned} \tag{10}$$

The new g_{best_new} provides a momentum in the velocity of the particles. The position of the particles consequently changes. So, the stagnancy in the movement of the particles can be overcome. To establish the effectiveness of blended Laplacian operator, few plots are provided in Fig. 3. The experimental analysis is performed for gastric cancer data, and the fitness values of searching particles for the two objectives t-score and accuracy are plotted for different iterations. After a 100 iteration when

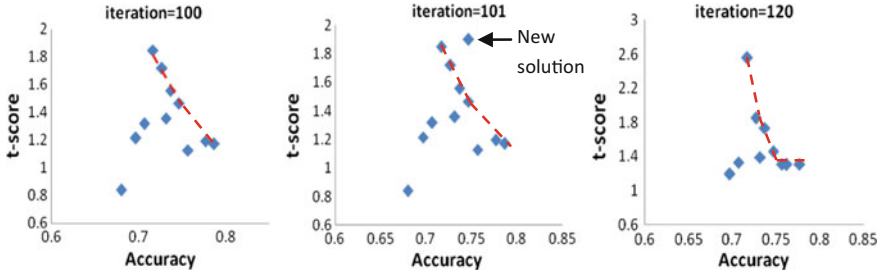

Fig. 3 Plot of fitness values for gastric cancer at different iterations

swam is unable to generate new better solution, a new solution is produced using blended operator. As a result, the swarm updates themselves and overcomes the stagnancy. The effect is shown for iteration number 120. New better solutions are generated by the searching particles. The overall MOBPSO technique is described in Algorithm 2.

Algorithm-2 (MOBPSO)

1. **Initialization.**
   Total Number of particle=N
   (a) Randomly initialize the position of particles, X_i (1=1, 2,…, N)
   (b)Initialize archive1 with few randomly chosen solutions
2. **Termination check.**
   (a) If the termination criterion holds go to step 8.
   (b) Else go to step 3.
3. **Set** t=1(iteration counter)
   For i= 1,2…N **Do**
   (a) **If** stagnancy occurs,
   Choose g_{best} randomly from the archive1
   Else choose g_{best} randomly from the archive2
   End If
   (b) Update the position according to Equations (8),(9)
   (c) Evaluate the fitness of the i^{th} particle $f_1(X_i)$ and $f_2(X_i)$ for two objectives
   End For
   (d) Choose the non-dominated solutions among N particles
   (e) Update the archive1 with non-dominated solutions
   (f) Check for stagnancy
   If stagnancy occurs
   i) Generate few new solutions (g_{best_new}) using blended laplacian operator as equation (10)
   ii) Construct a new archive2 using those g_{best_new}
   End If
4. **Set** t=t+1.
5. **Go to** step 2
6. Solution is the solution from archive1

4.2 Other Comparative Methods for the Selection of Genes

Most of the evolutionary and swarm intelligence algorithms such as genetic algorithm (GA), differential evolution (DE) and artificial bee colony (ABC) suffer from local trapping which results in premature convergence. The Laplacian blended operator can be implemented when such situation occurs as it produces few new solutions blending some previously generated solutions. So, the use of operator makes the optimization algorithms more efficient in the process of stochastic searching. Here, authors have integrated the blended operator with the above-mentioned algorithms and introduced multi-objective blended differential evolution (MOBDE), multi-objective blended artificial bee colony (MOBABC) and multi-objective blended genetic algorithm (MOBGA) in the similar fashion as it is done for MOBPSO. MOBDE, MOBABA and MOBGA are now applied to the four cancer datasets for marker gene selection. In the next subsection, the methodologies are discussed in brief.

4.2.1 Multi-Objective Blended Genetic Algorithm (MOBGA)

GA is a metaheuristic algorithm which is being inspired by the natural selection. It constitutes of few steps like parent selection, crossover and mutation (Yang and Honavar 1998). Initially, the algorithm starts with few solution termed as chromosome. Now, fitted chromosomes are considered as parents who are used to generate new child solutions. To create new solutions, a set of genetic operations like crossover and mutation are used. In MOBGA, initially non-dominated solutions are stored in an archive. Parents are chosen randomly from the archive to create next generation of solutions. Next, based on Pareto-optimal concept, fitted chromosomes survive and the archive is updated accordingly. Similar to the MOBPSO, when stagnancy occurs, blended Laplacian operator is utilized to overcome it. New parents are generated using blended Laplacian operator. For MOBGA, the process of gene selection is kept same as shown in Fig. 1, and only the MOBPSO block is replaced by the MOBGA.

4.2.2 Multi-Objective Blended Differential Evolution (MOBDE)

DE is a population-based stochastic optimization technique which adopts mutation and crossover operators to search for new promising areas in the search space (Price et al. 2006). The algorithm starts with a number of solutions based on non-dominated sorting and more promising solutions are kept in an archive. From the archive, fitted solutions are selected for mutation purpose and new solutions are produced. Similar to previously mentioned algorithms, when no further improvement is found, Laplacian blended operator is used. The binary format is implemented as done using Algorithm-1, and the process of gene selection is same as described in Fig. 1 except that the MOBPSO block is replaced by MOBDE.

Table 1 Parameters used in different swarm and evolutionary algorithms

Algorithms	Parameters	Explanation	Value
MOBPSO	N	Number of particle(s) in one swarm	20
	c_1, c_2	Acceleration constants	1.49
	w	Inertia	0.7
	r_1, r_2	Random numbers	[0, 1]
MOBGA	N	Number of genetic(s) in one group	20
	Ps	Selection ratio	0.8
	Pc	Crossover ratio	0.9
	Pm	Mutation ratio	0.01
MOBDE	N	Number of individual(s) in one group	20
	fm	Mutation factor	0.6
	CR	Crossover rate	0.9
MOBABC	N	Number of bee(s) in one swarm	20
	L	Limit for scout phase	100

4.2.3 Multi-Objective Blended Artificial Bee Colony (MOBABC)

Artificial bee colony (ABC) algorithm is inspired by the foraging behaviour of honey bees (Karaboga and Basturk 2007). Three groups of bees, employee bees, onlooker bees and scout bees, are involved in the searching process. The employee bee produces a modification on the position (solution) and depending on the non-dominated sorting procedure best positions are memorized. Here, those positions are stored in an archive. Onlooker bee chooses a food source from the archive and searches thoroughly across it. When stagnancy occurs, the archive is updated with new solutions, produced through Laplacian blended operator. MOBABC is applied to cancer datasets similar to the process as described in Fig. 1 just replacing the block of MOBPSO by MOBABC. The parameter settings of all other stochastic algorithms are given in Table 1.

5 Experimental Results

The experimental datasets consist of microarray data of Child_ALL, leukemia, colon and gastric cancer. Multi-objective blended GA (MOBGA), multi-objective blended DE (MOBDE), multi-objective blended ABC (MOBABC) and multi-objective blended PSO (MOBPSO) are employed for the task of feature gene selection using

supervised learning process in the field of cancer classification. The performance of the proposed multi-objective gene selection techniques is analysed and compared for four real-life cancers. The evaluation is performed based on classification results such as sensitivity, specificity, accuracy and F-score (Agarwalla and Mukhopadhyay 2016) using 10-fold cross-validation. Different classifiers are involved for the classification such as support vector machine (SVM), decision tree (DT), K-nearest neighbour (KNN) classifier and naive Bayes (NB) classifier (Kotsiantis et al. 2007). Experiments are carried out 10 times, and the average results are reported. The performance of MOBPSO is given in the subsequent sections utilizing different classifiers. Then, a comparative study is performed involving all the algorithms. Next, the results are compared with other existing methods reported in different research articles. The proposed methodologies establish promising results, indicating the capability to produce more effective gene selection activity.

5.1 Classification Results

In this chapter, the aim is to identify top differentially expressed genes (DEGs) which are performing well in the process of classification. By involving MOBPSO, MOBGA, MOBDE and MOBABC algorithms, the optimized gene subset is obtained. The gene subset which is identified is validated by analysing the classification results. The proposed methodology is implemented using four different well-known classifiers (SVM, DT, KNN and NB). The experimental result ascertains that the proposed methodology is able to extract important features from the huge dataset. The classification results of MOBPSO algorithm for different cancer datasets are given in Table 2. For leukemia cancer, NB classifier shows better performance compared to others classifiers. Here, 100% accuracy is achieved which indicates the perfect classification of disease. For colon cancer, decision tree classifier is working efficiently in terms of providing good specificity of the result. Highest accuracy is achieved by the SVM classifier which is equal to 87%.

For gastric cancer, SVM achieves 89% accuracy which establishes its superiority over the other classifiers, used for the experiment. KNN classifier is providing 79% accuracy and 89% sensitivity as the classification result of Child_ALL data.

The accuracy of classification obtained using different classifiers is also given in the form of bar chart in Fig. 4 for better interpretability of the results. The comparative result shows that for leukemia data, NB and decision tree both work effectively to classify the cancer. In case of colon cancer, SVM classifier is producing reliable result. For gastric cancer, SVM and NB classifiers are able to find out relevant genes for disease classification. KNN classifier is performing top for Child_ALL cancer compared to all other classifier techniques. Similar to MOBPSO, the other methodologies like MOBGA, MOBABC and MOBDE are applied on the cancer datasets and a comparative study is performed in the next subsection.

Table 2 Results of classification using MOBPSO for different cancer datasets

Dataset	Algorithms	Sensitivity	Specificity	Accuracy	F-score
Leukemia	*MOBPSO*-SVM	1.00	0.78	0.89	0.87
	MOBPSO-KNN	0.89	0.87	0.89	0.91
	MOBPSO-tree	0.96	1.00	0.99	0.98
	MOBPSO-NB	**1.00**	**1.00**	**1.00**	**1.00**
Colon	*MOBPSO*-SVM	**0.91**	0.72	**0.87**	**0.81**
	MOBPSO-KNN	0.75	0.71	0.73	0.70
	MOBPSO-tree	0.72	**0.77**	0.77	0.75
	MOBPSO-NB	0.78	0.76	0.76	0.75
Gastric	*MOBPSO*-SVM	**1.00**	**0.89**	**0.89**	**0.92**
	MOBPSO-KNN	0.81	0.87	0.83	0.89
	MOBPSO-tree	0.78	0.83	0.82	0.86
	MOBPSO-NB	0.91	0.87	0.89	0.90
Child_ALL	*MOBPSO*-SVM	0.71	0.76	0.70	0.71
	MOBPSO-KNN	**0.89**	**0.73**	**0.78**	**0.81**
	MOBPSO-tree	0.78	0.77	0.74	0.72
	MOBPSO-NB	0.72	0.73	0.71	0.69

5.2 Comparative Analysis

To estimate the effectiveness of the proposed method, experiments are conducted on the four real-time cancer datasets. Here, authors have provided the results of classification of disease after applying MOBPSO, MOBDE, MOBGA and MOBABC on the datasets. SVM classifier is used for each classification purpose. Average result of 10 times 10-fold cross-validation is reported for the comparative study in Table 3. Best results are marked in bold. For leukemia, good results are obtained using MOBGA. For colon cancer, MOBPSO is the best performing feature selection technique and MOBABC is able to obtain second position. MOBDE has achieved promising result for gastric cancer, whereas all the algorithms are able to achieve 100% sensitivity for the data. For Child_ALL data, MOBPSO is able to estimate the proper genes for the classification of disease with an accuracy of 75%. The

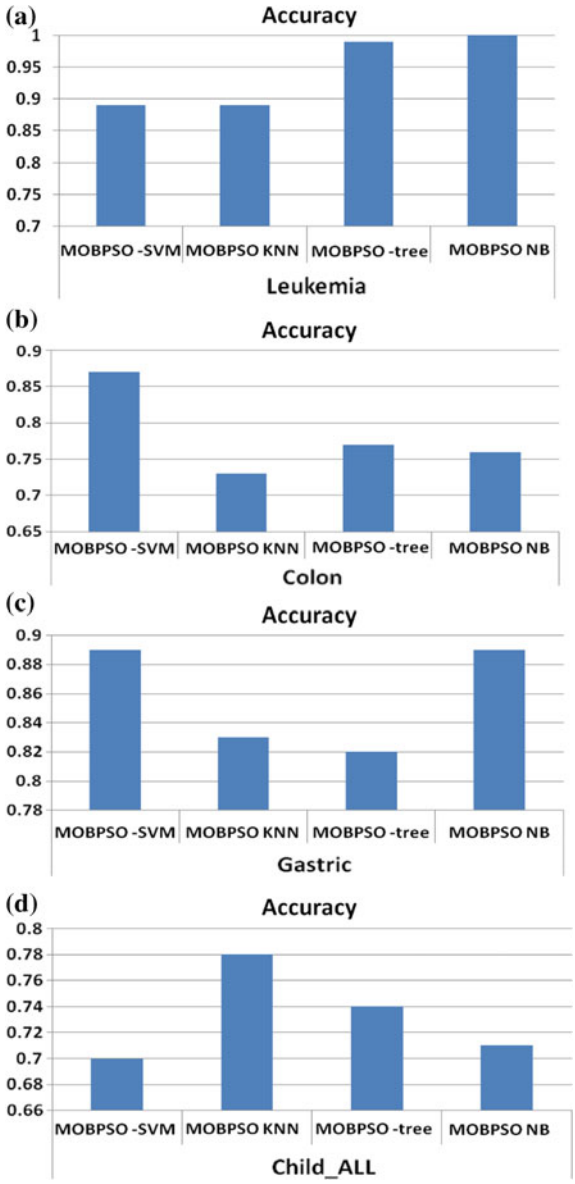

Fig. 4 Accuracy of classification using different classifiers for **a** leukemia, **b** colon, **c** gastric, **d** Child_ALL cancer datasets

comparison of accuracy obtained from different proposed algorithms is also shown in Fig. 5.

In Table 4, results are again compared with other approaches, reported in different literature for gene selection methodology (Mukhopadhyay and Mandal 2014; Apolloni et al. 2016; Salem et al. 2017; Luo et al. 2011). NSGA-II (Deb et al. 2002) and

Table 3 Result of comparison with different swarm and evolutionary algorithms

Dataset	Algorithms	Sensitivity	Specificity	Accuracy	F-score
Leukemia	*MOBPSO*	**1.00**	0.78	0.89	0.87
	MOBABC	0.84	0.86	0.83	0.80
	MOBDE	0.71	0.83	0.81	0.79
	MOBGA	0.90	**0.91**	**0.90**	**0.89**
Colon	*MOBPSO*	**0.91**	0.72	**0.87**	**0.81**
	MOBABC	0.90	**0.76**	0.81	0.83
	MOBDE	0.71	0.69	0.67	0.68
	MOBGA	0.78	0.73	0.77	0.80
Gastric	*MOBPSO*	**1.00**	0.89	0.89	0.92
	MOBABC	**1.00**	0.90	0.90	0.92
	MOBDE	**1.00**	**0.94**	**0.91**	**0.93**
	MOBGA	**1.00**	0.86	0.87	0.89
Child _ALL	*MOBPSO*	**0.71**	**0.76**	**0.75**	**0.71**
	MOBABC	0.60	0.64	0.61	0.62
	MOBDE	0.65	0.70	0.67	0.63
	MOBGA	0.50	0.70	0.66	0.68

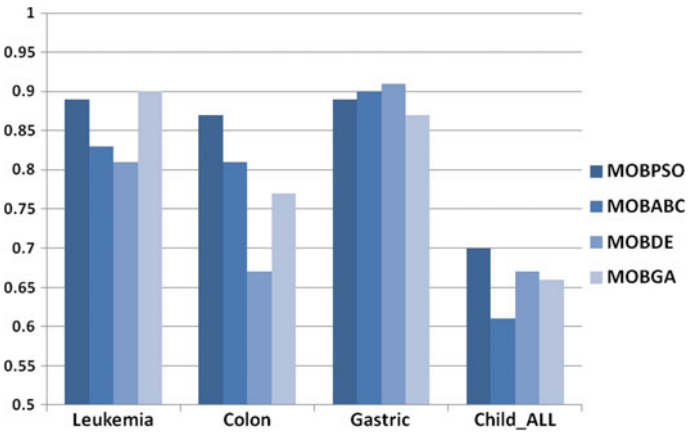

Fig. 5 Accuracy of classification using different algorithms

MOEA/D (Zhang and Li 2007) are also applied on the cancer datasets to obtain Pareto solutions for the objectives. For colon and Child_ALL datasets, MOBPSO is able to gain the best result of accuracy in classification of cancer among the techniques, used for comparison. For other two datasets, the results are also quite promising. The comparative result signifies the efficiency of the proposed methodology for supervised cancer classification.

Table 4 Comparison of accuracy of classification with other results reported in the literature

Reference	Year	Leukemia	Colon	Gastric	Child_ALL
(Salem et al. 2017)	2017	0.97	0.85	–	–
(Luo et al. 2011)	2011	0.71	0.80	–	–
(Apolloni et al. 2016)	2016	0.82	0.75	–	–
(Mukhopadhyay and Mandal 2014)	2014	–	–	0.96	0.74
NSGA-II (Deb et al. 2002)	2002	0.78	0.75	0.93	0.68
MOEA/D (Zhang and Li 2007)	2007	0.92	0.81	0.91	0.72
MOBPSO	2017	0.89	0.87	0.89	0.75
MOBDE	2017	0.83	0.81	0.90	0.61
MOBGA	2017	0.81	0.67	0.91	0.67
MOBABC	2017	0.90	0.77	0.87	0.66

Table 5 Biological significance for gene–disease association

Dataset	Associated diseases	Gene symbol
Leukemia	Leukemia	RAG1(3), MSH(61), CD36(2)
	Lymphomas	CCND3(7), LYN(4)
Colon	Colorectal cancer	MAPK3(11), EGR1(1)
	Malignant tumour of colon	IGF1(67), KLK3(781)
Gastric	Malignant neoplasm of stomach	CYP2C9(1), SPP1(20)
	Stomach carcinoma	SPP1(21), NOS2(2),
Child_ALL	Tumour progression	SMAD3(1), ITGA6(1)

5.3 Biological Relevance

Biological relevance of the experimentally selected genes is analysed by gathering the information about those genes from disease–gene association database. Also, the information of number of Pubmed citations against those genes is collected. In Table 5, disease information related to those top genes is given. For example, MSH gene has 61 Pubmed citations as evidence that the gene is related to leukemia cancer. Similarly, for colon cancer KLK3 is the most cited gene related to the disease. The information proves the biological significance of the proposed work. As a whole, it can be concluded that the proposed gene selection methodologies are more efficient in detection of the relevant genes for all the different types of datasets.

6 Conclusion

Classification of disease through supervised learning method leads to the investigation on feature selection technique. So, for the feature reduction and extraction from the huge dimension of data, authors involve new multi-objective blended particle swarm optimization (MOBPSO) technique. The methodology uses a new concept of integrating blended Laplacian operator in the algorithmic portion, and it generates a subset of genes based on two objectives. The multi-objective concept along with the proposed methodology is proved to be very useful in the context of diagnosis of disease as it identifies biologically significant genes related to the disease. Similarly, authors have implemented the concept with other swarm and evolutionary algorithms and developed multi-objective blended differential evolution (MOBDE), multi-objective blended artificial bee colony (MOBABC) and multi-objective blended genetic algorithm (MOBGA). The experimental result establishes that the proposed technique is able to provide promising result in the context of classification of disease which reflects its effectiveness of selecting relevant feature genes.

References

P. Agarwalla, S. Mukhopadhyay, Selection of relevant genes for pediatric leukemia using co-operative Multiswarm. Mater. Today Proc. **3**(10), 3328–3336 (2016)

U. Alon, N. Barkai, D.A. Notterman, K. Gish, S. Ybarra, D. Mack, A.J. Levine, Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc. Natl. Acad. Sci. **96**(12), 6745–6750 (1999)

J. Apolloni, G. Leguizamón, E. Alba, Two hybrid wrapper-filter feature selection algorithms applied to high-dimensional microarray experiments. Appl. Soft Comput. **38**, 922–932 (2016)

S. Bandyopadhyay, S. Mallik, A. Mukhopadhyay, A survey and comparative study of statistical tests for identifying differential expression from microarray data. IEEE/ACM Trans. Comput. Biol. Bioinform. **11**(1), 95–115 (2014)

K.H. Chen, K.J. Wang, M.L. Tsai, K.M. Wang, A.M. Adrian, W.C. Cheng, K.S. Chang, Gene selection for cancer identification: a decision tree model empowered by particle swarm optimization algorithm. BMC Bioinform. **15**(1), 49 (2014)

M.H. Cheok, W. Yang, C.H. Pui, J.R. Downing, C. Cheng, C.W. Naeve, W.E. Evans, Treatment-specific changes in gene expression discriminate in vivo drug response in human leukemia cells. Nat. Genet. **34**(1), 85–90 (2003)

K. Deb, A. Pratap, S. Agarwal, T.A.M.T. Meyarivan, A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE Trans. Evol. Comput. **6**(2), 182–197 (2002)

T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, C.D. Bloomfield, Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science **286**(5439), 531–537 (1999)

A.O. Hero, G. Fluery, Pareto-optimal methods for gene filtering. J. Am. Stat. Assoc. (JASA) (2002)

Y. Hippo, H. Taniguchi, S. Tsutsumi, N. Machida, J.M. Chong, M. Fukayama, H. Aburatani, Global gene expression analysis of gastric cancer by oligonucleotide microarrays. Cancer Res. **62**(1), 233–240 (2002)

D. Karaboga, B. Basturk, A powerful and efficient algorithm for numerical function optimization: artificial bee colony (ABC) algorithm. J. Glob. Optim. **39**(3), 459–471 (2007)

J. Kennedy, Particle swarm optimization. *Encyclopedia of Machine Learning* (Springer, US, 2011), pp. 760–766

N. Khunlertgit, B.J. Yoon, Identification of robust pathway markers for cancer through rank-based pathway activity inference. Adv. Bioinform. (2013)

S.B. Kotsiantis, I. Zaharakis, P. Pintelas, Supervised machine learning: a review of classification techniques (2007)

L.K. Luo, D.F. Huang, L.J. Ye, Q.F. Zhou, G.F. Shao, H. Peng, Improving the computational efficiency of recursive cluster elimination for gene selection. IEEE/ACM Trans. Comput. Biol. Bioinform. **8**(1), 122–129 (2011)

M.S. Mohamad, S. Omatu, S. Deris, M.F. Misman, M. Yoshioka, A multi-objective strategy in genetic algorithms for gene selection of gene expression data. Artif. Life Robot. **13**(2), 410–413 (2009)

A. Mukhopadhyay, M. Mandal, Identifying non-redundant gene markers from microarray data: a multiobjective variable length PSO-based approach. IEEE/ACM Trans. Comput. Biol. Bioinform. (TCBB) **11**(6), 1170–1183 (2014)

K. Price, R.M. Storn, J.A. Lampinen, *Differential Evolution: A Practical Approach to Global Optimization* (Springer Science & Business Media, 2006)

H. Salem, G. Attiya, N. El-Fishawy, Classification of human cancer diseases by gene expression profiles. Appl. Soft Comput. **50**, 124–134 (2017)

N. Srinivas, K. Deb, Muiltiobjective optimization using nondominated sorting in genetic algorithms. Evol. Comput. **2**(3), 221–248 (1994)

A.V. Ushakov, X. Klimentova, I. Vasilyev, Bi-level and bi-objective p-median type problems for integrative clustering: application to analysis of cancer gene-expression and drug-response data. IEEE/ACM Trans. Comput. Biol. Bioinform. (2016)

J. Yang, V. Honavar, Feature subset selection using a genetic algorithm. IEEE Intell. Syst. Appl. **13**(2), 44–49 (1998)

L. Zhang, J. Kuljis, X. Liu, Information visualization for DNA microarray data analysis: a critical review. IEEE Trans. Syst. Man Cybern. Part C Appl. Rev. **38**(1), 42–54 (2008)

Q. Zhang, H. Li, MOEA/D: a multiobjective evolutionary algorithm based on decomposition. IEEE Trans. Evol. Comput. **11**(6), 712–731 (2007)

C.H. Zheng, W. Yang, Y.W. Chong, J.F. Xia, Identification of mutated driver pathways in cancer using a multi-objective optimization model. Comput. Biol. Med. **72**, 22–29 (2016)