Genomics 110 (2018) 10–17

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Gene selection for tumor classification using a novel bio-inspired multi-objective approach

M. Dashtban[a,*], Mohammadali Balafar[a], Prashanth Suravajhala[b,c]

[a] Department of Computer Engineering, Faculty of Electrical & Computer Engineering, University of Tabriz, Iran
[b] Birla Institute of Scientific Research, Statue Circle, Jaipur 302001, Rajasthan, India
[c] Bioclues.org, Kukatpally, Hyderabad 500072, Telangana, India

ARTICLE INFO

Keywords:
Cancer classification
Gene selection
Bat algorithm
Feature selection
Evolutionary algorithms
Microarray data analysis

ABSTRACT

Identifying the informative genes has always been a major step in microarray data analysis. The complexity of various cancer datasets makes this issue still challenging. In this paper, a novel Bio-inspired Multi-objective algorithm is proposed for gene selection in microarray data classification specifically in the binary domain of feature selection. The presented method extends the traditional Bat Algorithm with refined formulations, effective multi-objective operators, and novel local search strategies employing social learning concepts in designing random walks. A hybrid model using the Fisher criterion is then applied to three widely-used microarray cancer datasets to explore significant biomarkers which reveal the effectiveness of the proposed method for genomic analysis. Experimental results unveil new combinations of informative biomarkers have association with other studies.

1. Introduction

Machine learning has been an extremely powerful tool for biological data analysis. It has had several applications in various fields of biological sciences mostly in two recent decades. Designing the prediction models is one of the most interesting applications of machine learning. Prediction models have been used in different biological applications such as [1–12]. Developing statistical models for identification and classification of cancerous tissues from normal tissues using gene expression profiles is one the most challenging application of ML [13]. Novel DNA microarray technology has the feasibility of measuring the expression levels of huge number of genes simultaneously in a single experiment. This technology enables the researchers to comprehensive overview to precisely discover which genes are expressed in a specific tissue under various conditions. However, developing prediction models using gene expression profiles is quite challenging since there are several irrelevant or insignificant genes to clinical diagnosis and prognosis [14]. Hence, identifying highly informative genes for cancer classification is accordingly an valuable endeavor.

Gene selection is a branch of feature selection that is the process of selecting the subsets of relevant and significant genes for a classification/prediction problem. Several issues are associated with gene selection in microarray datasets [15]. First, selecting a subset of informative genes from high-dimensional microarray datasets is a non-deterministic polynomial-time (NP)-Hard problem. Therefore, meta-heuristics algorithms including evolutionary methods or Bio-inspired algorithms are widely used. The next issue is technically the curse of sparsity, which means the number of samples is very small and scanty. Next, the high complexity of gene expression data that arises from several facts such as the high correlation between genes and considerable interactions among them, makes the process of selecting informative genes very challenging. For instance, a high regulated. Furthermore, the diagnostic process of identifying infected tissues would be quite easier, reliable and interpretable when the number of informative genes is small [51-53].

Several gene selection methods have been proposed in the literature to surmount the challenges that roughly fall into three categories including filter model, wrapper model, and hybrid model [16]. A filter model relies primarily on the general statistical characteristics of the training data without using any learning algorithm. Thus, such methods are fast but have rather poor performance. In contrast, the wrapper models use a predetermined learning algorithm to guide the searching process toward optimal subset(s) of features. This model often employs bio-inspired or evolutionary algorithms in its body so that it starts with a population of features subsets. Such population should be evaluated using that established learner and be improved in several iterations. Hence, the computational complexity of this approach is high especially in the case of high-dimensional datasets [17–19]. The performance of

* Corresponding author.
E-mail address: dashtban@tabrizu.ac.ir (M. Dashtban).

http://dx.doi.org/10.1016/j.ygeno.2017.07.010
Received 19 March 2017; Received in revised form 12 July 2017; Accepted 30 July 2017
Available online 03 August 2017

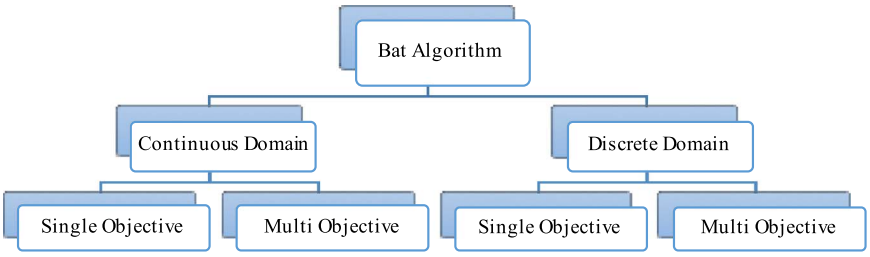


Fig. 1. A general classification of bat algorithm-based approaches proposed in the literature according to encoding representation and number of objectives.

Pseudocode 1
MOBBA-LS: multi-objective binary bat algorithm with local searches.

INPUT: population size M, number of features n,
Stopping Criterion, loudness A, pulse emission rate r
injectRate iR, walkRate wR, Adjustment rates α, γ, σ
OUTPUT: Subset of features in the latest population
For each bat bi ($\forall i$=1,...,m), do
  For each feature j ($\forall j$=1,...,n), do
    $x_i^j \leftarrow$Random{True, False}
  End
  $v_i^j \leftarrow$ min velocity
  $r_i \leftarrow$ *Random* [min pulse rate, max pulse rate]
  $cost_i \leftarrow$ [evaluate bat bi, #genes in bi]
  $t \leftarrow$ initial zero iteration
  $x_t^j \leftarrow$Random{0,1}
End
$t \leftarrow first\ iteration$
$A_t \leftarrow$ max loudness
Initial population P_t using x, v, r, and Costs
$\boldsymbol{F1} \leftarrow fast-non-dominated-sorting(P_t)$
While *stopping criteria met* do
  For each bat Bi ($\forall i$=1,...,M)
    Generate new Bat $\mathcal{N}$Bi using the Eq. (2), (3), & (7)
    Evaluate Bi using an evaluation function
    If ($rand < A_t$ And $\mathcal{N}$Bi dominate B_i) then
      Update the Pulse rate using Eq. (5)
    End If
  End
  $\boldsymbol{F1} \leftarrow fast-non-dominated-sorting(P_t)$
  Random Walk using local search strategies
  $t \leftarrow next\ iteration$
  Update the loudness using Eq. (6)
End While
Return P_t

wrapper approaches is generally better than the filter models since they consider interactions between the solutions and the predictors. The hybrid approaches take the advantages of both models by firstly applying a filter model to reduce the feature space and selecting more relevant genes, and then exploiting a wrapper model to search for optimal subset/s of features [20–22].

Furthermore, several nature-inspired algorithms have been developed in the literature in recent decades [23,24]. Bat algorithm (BA) [25] was one of the most recent Bio-inspired methods that soon became well known for its superior performance in solving several optimization problems. BA was widely used mainly for its faster convergence and better time complexity. Several Bat-inspired methods exist in the literature, but, to our knowledge, most of them are either mono-objective or designed for solving optimization problems in the continuous domain. In this work, we developed a multi-objective version of the BA with refined formulations, effective multi-objective operators and robust local search strategies particularly for variable selection in binary domain namely MOBBA-LS. The main contributions of the proposed method are as follows:

- The first ever multi-objective version of bat algorithm for binary variable selection,
- Novel local search strategies incorporating social learning concepts
- Specific random walk well-suited for local search in binary domain
- Simplified multi-objective procedure without generating multiple population and its associated computational burden.

The proposed method was then applied on three high-dimensional microarray cancer datasets to identify significant biomarkers and to investigate the effectiveness and the usefulness of MOBBA-LS for genomic analysis. It found new combinations of the most discriminative biomarkers competitive with previous studies.

2. Materials and methods

A hybrid model was applied to microarray datasets for gene selection in tumor classification. A filter method was first used to efficiently screen out the highly irrelevant genes and form a filtered subset of a relatively small size. For initial filtering, the Fisher score that have a proven performance in separating informative genes [26,27] and in other applications [28,29] was utilized. Fisher criterion employs the statistical properties of each gene in different classes as a potential measure of discriminant ability for classification (see Eq. (1)). In this study, only 500 top-scored genes (as suggested by [26,30]) were chosen as the filtered subset.

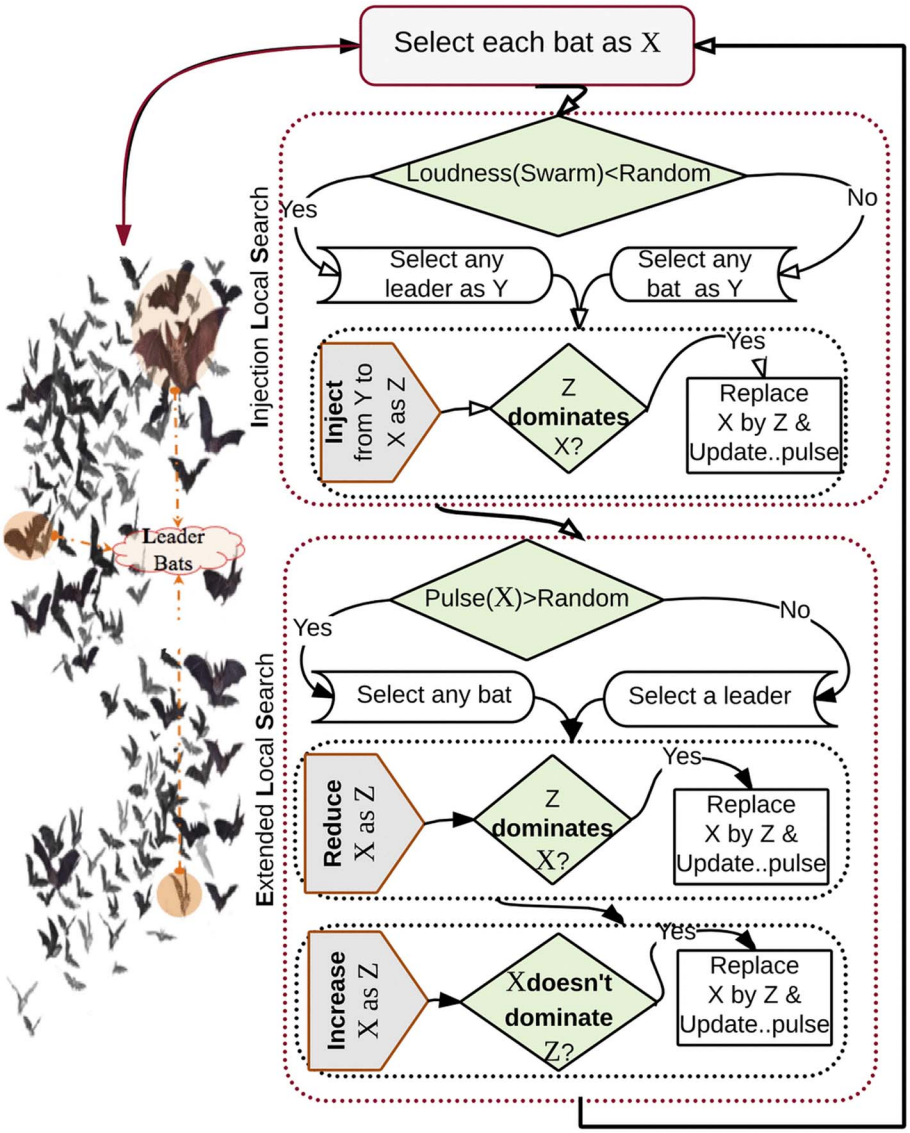
$$FisherSc(\mathrm{t}) = \sum_{j=1}^{M} (\mu_{ij} - \mu_i)^2 \Big/ \sigma_{ij}^2 \tag{1}$$

where, M is the number of classes, μ_i is the arithmetic mean of ith gene expression values, μ_{ij} is the arithmetic mean of values of jth class of ith gene and σ_{ij}^2 is the variance of expression values of ith gene in its jth class.

The filtered subset was then employed by a new-developed wrapper algorithm to select the final subsets of highly discriminating genes for cancer classification. The developed wrapper method is indeed a novel extension of bat algorithm. Several versions of BA have been proposed in the literature for various application domain. Those algorithms could be typically classified in terms of representation types and how the algorithm deal with the problem objectives as illustrated in Fig. 1. The BA has mostly adopted in continuous domain for solving engineering optimization problems [24,31]. Nonetheless, there are a few number of Binary BA exist in the literature all of which, to best of our knowledge, are single objective i.e., they did not solve the variable selection problem as a multi-objective optimization problem such as [18,32–34]. In contrast, in this study, a multi-objective version of Binary BA integrating various artificial intelligence concepts was proposed.

2.1. Bat algorithm

BA is a natural-inspired algorithm that was computerized by employing the fundamental characteristics of microbats in finding their preys as described in [25]. The preys in computerized version are



Flowchart 1. Bio-inspired local search strategies.

Table 1
Identified genes in Leukemia cancer dataset and number of misclassified samples of four classifiers.

Data	Leukemia				Genes		
	Classifiers				Accession code	Description	Database-Id
	SVM	KNN	NBY	DT			
Train	0	0	0	0	M92287_at	CCND3 Cyclin	2354
Test	1	0	0	3	X95735_at	D3	4847
LOOCV	1	1	0	1	HG1612-HT1612_at	Zyxin Macmarcks	804

DT stands for Decision Tree classifier, NBY for Naïve Bayes, KNN for K-nearest-neighbor.

indeed the optimal solutions being sought. The bats, in nature, fly randomly with different velocity v_i from a position x_i with frequency f_{min} varying wavelength λ and loudness A to search for prey, as formulated in Eqs. (2)–(5).

$$f_i = (f_{max} - f_{min})\beta \tag{2}$$

$$v_i^t = v_i^{t-1} + (x_i^{t-1} - x_*)f_i \tag{3}$$

$$x_i^j = x_i^{t-1} + v_i^t \tag{4}$$

$$A_i^{t+1} = \alpha A_i^t,\ r_i^{t+1} = r_i^0[-\gamma t] \tag{5}$$

where α and γ is a real number between 0.0 and 1.0. In this study, the Loudness of singular bats (A_i^{t+1}) was redefined as a total swarm's loudness in every iteration as shown in Eq. (6). The total loudness brings the possibility of maintaining the diversity and exploitation of the swarm in an easier way and with lower space complexity particularly in the local search level.

$$A^{t+1} = \alpha A^t \tag{6}$$

Moreover, among a few representative binary BA in the literature [18,32,33], Nakamura used a sigmoid function to map the continuous parameters of the BA into the binary values nonlinearly. We preferred that non-linear transformation since that tackled out-of-bound velocity values very well (see Eq. (7)).

$$x_i^t = \begin{cases} 1 & if\ \ S(v_i^t) > \sigma \\ 0 & otherwise \end{cases}, \quad S(v_i^t) = \frac{1}{1 + e^{-v_i^t}} \tag{7}$$

2.2. Multi-objective binary bat algorithm with specific local searches

A new wrapper method exploiting the refined Bat algorithm, fast multi-objective evolutionary operators, and novel specific local search strategies, namely MOBBA-LS, was proposed. The MOBBA-LS was

Table 2
The misclassification counts of the identified genes without including Clone-ID HG1612-HT1612_at.

M92287_at, X95735_at	Classifiers			
	SVM	KNN	NBY	DT
Train	0	0	1	0
Test	2	2	2	3
LOOCV	2	2	3	2

Table 3
Identified genes in Prostate cancer dataset and number of misclassified samples of four classifiers.

Data	Prostate				Genes	
	Classifiers				Accession	Index
	SVM	KNN	NBY	DT	37639_at, 37939_at	6185, 6247
Train	0	4	9	1	40607_at	9937
Test	2	1	1	25	41504_s_at	10234
LOOCV	8	7	10	8	38091_at, 38044_at	9097, 9050

DT stands for Decision Tree classifier, NBY for Naïve Bayes, KNN for K-nearest-neighbor.

Table 4
Identified genes in SRBCT cancer dataset and number of misclassified samples of four classifiers.

Data	SRBCT				Genes	
	Classifiers				Image-Id	Database-Id
	SVM	KNN	NBY	DT	1435862, 461425	545, 554
Train	0	0	3	1	812105, 755239	742, 801
Test	3	0	0	4	842973, 244618	1666, 2046
LOOCV	7	6	3	7		

developed with binary representation to be well suited for variable selection problems. In the view of multi-objective behaviour, MOBBA-LS takes the advantages of multi-objective operators of NSGA-II [35] that is one of the most efficient and fast multi-objective algorithms in the literature. Nevertheless, unlike NSGA-II, the MOBBA-LS does not generate multiple population and therefore reduce its related computational burdens; instead it focuses on altering and manipulating the current swarm wisely to increase diversity while maintaining exploitation. The MOBBA-LS adopts the *fast-non-dominated-sort* operator to identify the leader bats that are our superior solutions and technically are the members of the first front of multi-objective output. This leader set is *then* effectively exploited by local searches. The notion of employing a leader set was inspired by social learning strategies [36] and was recently employed in designing a modern particle swarm optimization algorithm to improve its local search [37]. The leader bats were used in the local search procedures to guide the swarm smoothly toward the potential sources (the search regions). Two novel local search strategies were adopted to further the search potential of the proposed algorithm for variable selection specifically in binary domain. Both methods endeavor wisely to maintain exploitation and exploration, for instance, by randomizing the selection process among leader bats and other bats, and by intelligent movements in the search space.

The general steps of the proposed algorithm are demonstrated in Pseudocode 1. From that, after some parameter initializations in Lines 1 to 3, random binary bats are created through Lines 5 to 17. The leader Bats are separated in Line 18. The main loop of the MOBBA-LS iterates through Lines 19 to 31. Generating new solutions based on Bat Motion rules is described in Lines 20-26. The prediction capability of each bat/solution was evaluated using a classifier. The choice of classifiers during metaheuristics' search is critical to properly guide the searching procedure and to obtain informative genes in microarray data classification (as discussed in our previous work [26]). In Line 22, SVM classifier with a 10-fold cross-validation was used to assess the classification value of each subset. In Line 27, the bats are non-dominantly sorted to obtain the first non-dominated front (the leader bats) to be used afterward by the local searches in the Line 28. The procedure iterates until the stopping criteria meet. Some stopping criteria were defined including (a) reaching to minimum prediction accuracy on training samples, (b) obtaining at least one solution with a minimum length among superior solutions, and (c) reaching to a predetermined maximum iteration.

2.2.1. Local search strategies

Local search strategies play an important role in every evolutionary algorithm. As a matter of fact, incorporating local search methods offers "not only a better speed of convergence to the evolutionary approach, but also better accuracy for the final solutions" [38]. In the presented algorithm, two novel intelligent local search strategies, namely, injection local search (ILS) and extended local search (ELS) were designed and adopted to improve the searching potential of MOBBA-LS. Flowchart 1 illustrates how two strategies work together. Each strategy involves selection process and random walk operations. The selection process focuses on whether the leaders or typical bats should be inspected over iterations. The Loudness of whole swarm and pulse rate of bats were exploited to obtain a trade-off in the selection process of ILS and ELS, respectively. In both strategies, the concentration on improving the leaders gradually increases as the Loudness or pulse rate approaches their boundary values.

The first strategy, the ILS, is to further the exploitation of sound characteristics of every bird in the swarm by smoothly distributing them via injection operator. The ILS actually imitates the social behavior of the bats to learn from each other along with learning from superior ones. The notion of this strategy is inspired from social learning concepts that have been successfully utilized in the modern extensions of particle swarm optimization [39]. In ILS, the position of a bat is subject to change based on either the position of a leader or a neighboring bat. The injection operator replaces some features within the position of the bat (as need-to-be-altered characteristics) with that of a randomly selected one (as learned characteristics). The amount of changes adjusts by a predetermined injection rate.

The second strategy incorporates exploitation while focusing on

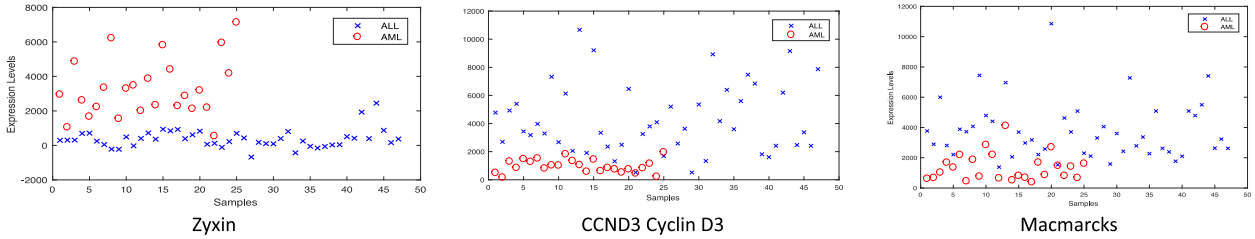


Fig. 2. Scatterplot of three identified genes in different samples of Leukemia cancer dataset. Zyxin is highly expressed for AML, CCND3 is highly expressed for all and Macmarcks is highly expressed for all and moderately for a few samples of AML.

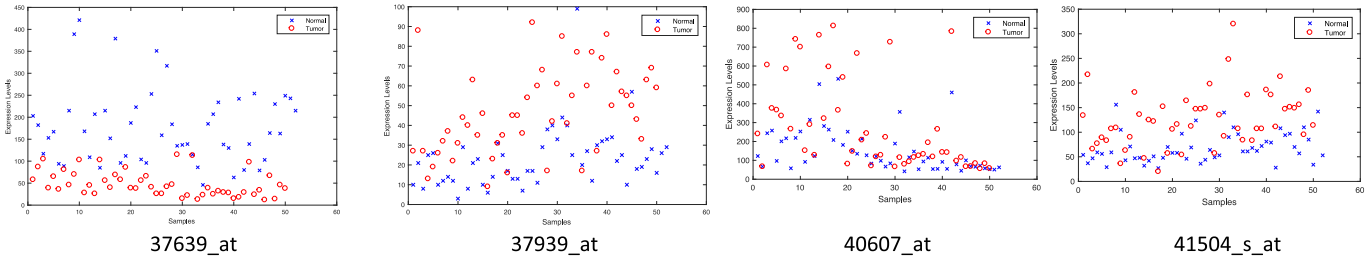


Fig. 3. Scatterplot of four identified genes in the training samples of prostate cancer dataset. The red and blue colors indicate the tumor and normal tissues, respectively. The image Id 37639_at is highly expressed for the normal tissues, 37939_at is moderately to highly expressed for tumor tissues and a few normal samples, the Gene ID 40607_at is highly expressed for tumor samples and moderately expressed for a few tumor samples and 41504_s_at is moderately to highly expressed for tumor samples.(For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

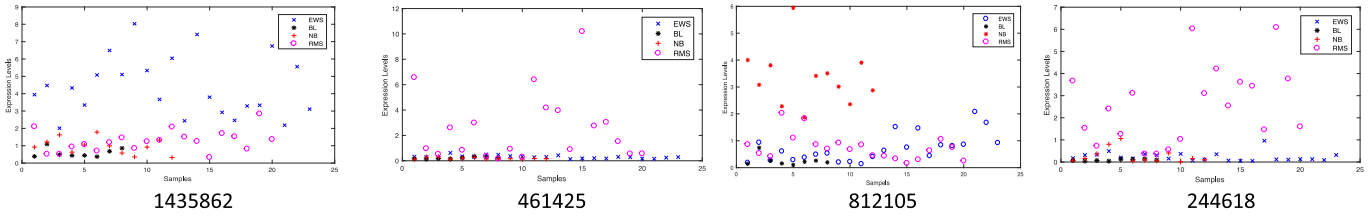


Fig. 4. Scatterplot of four identified genes in the training samples of SRBCT cancer dataset. The blue, black, red, and maroon colors correspond to EWS, BL, NB and RMS class of cancer, respectively. The Clone ID 1435862 is highly expressed for EWS, 461425 and 244618 are highly expressed for RMS, and 812105 is highly expressed for NB with a few cases moderately expressed for EWS and RMS.(For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

exploration. the ELS encourages each bat to perform some specific random walks via two specific operators namely Reduce Operator (RO) and Increase Operator (IO). The RO alters the position of the bat by removing randomly some features along its length whereas the IO arbitrarily augments some features. In both operators, the pulse rate (of the current bat) increases when a better position unfolds. All operations function smoothly at a predetermined tiny walk rate. One important consideration in the proposed search strategy is the possibility of *backward search* by creating and prioritizing the solution with more number of features by the IO operator. Such behaviour is clear in the last block of Flowchart 1 where the previously found solution X would be preserved only of it fully dominates the new solution that has slightly more number of features. The ELS could be seen as an intelligent type of mutation operator adopted as the random walk for the bat algorithm. One consideration is the walk rate and injection rate must be set to tiny values (such as 0.01) to avoid fluctuations around the optimum points of the search space. However, it can be empirically set to higher values to facilitate the convergence of the algorithm.

3. Experimental results

The proposed method was applied on three high-dimensional microarray cancer dataset to identify most discriminative biomarkers. First, the Fisher score was employed to filter the datasets and to create a new dataset with only top statistically relevant genes. The MOBBA-LS was then applied 30 times (as suggested by [17]) to obtain most discriminating genes in each dataset. Afterward, the identified subsets of genes in the first front of the proposed multi-objective algorithm were studied over independent runs. A 10-fold cross validation with SVM was employed to approximate the prediction accuracy of the solutions (the bats) during the algorithm process (see Supplementary materials 2 for parameter settings and implementations). The efficiency of utilizing the SVM for measuring the quality of population individuals within an evolutionary algorithm was already proved in microarray data application [26,40,41]. Tables 1–4 explain the best identified subsets with lowest number of genes and highest accuracy. The prediction accuracy of each subset were evaluated using four widely-used classifiers including support vector machines (SVM), K-Nearest Neighbors (KNN), Naïve Bayes (NBY), and Decision Tree (DT), each of which was assessed in according to the training samples, testing samples, and Leave-One-Out Cross Validation (LOOCV). All implementations were conducted in Matlab 2016b MacOS can also be accessed online in mathworks.com searching for MOBBA-LS method.

3.1. Leukemia cancer data

The first dataset we used was the leukemia cancer data containing 7129 gene-expression levels of 72 patients with either acute myeloid leukemia (AML) or acute lymphoblastic leukemia (ALL) [14]. This two-class dataset was originally divided into a training set of 38 samples and a test set of 34 samples. The training samples consisted of 27 ALL and 11 AML samples. The test data had 20 and 14 samples for ALL and AML, respectively.

The proposed method identified a highly informative subset of three genes which were mostly consistent with the previous studies in the literature. Table 1 presents a description of the best-identified subset of genes with its performance in various classifiers. From Table 1, all classifiers were suited well on training samples with zero misclassified samples. On testing samples, both SVM and Decision-Tree (DT) had only one miss in AML samples while DT had additionally two more misses in ALL samples. However, KNN and NBY performed best on both the training and testing samples with zero misclassification rate. The LOOCV performance of these genes was also remarkable. All classifiers had at most one misclassified sample out of 72 ones. Overall, the SVM, KNN and NBY had very close performance using these genes even though that NBY outperformed others with perfect prediction accuracy over all data subsets.

Furthermore, the biological relevance of these genes was also remarkable. The first two genes have already been identified as informative genes by [15,42]. Also, several combinations of M92287_at were determined by [43] using rough sets even so none of them gave the overall accuracy of more than 94.1%. Our results suggest that the combination of Macmarcks could be quite effective in prognosis of leukemia cancer.

3.1.1. The combination of identified genes without Macmarcks

From Table 2 the number of misclassified samples in test data and LOOCV samples were dramatically increased while the same

Table 5
Comparing the performance of MOBBA-LS with the literature methods.

Dataset	Method	Accuracy	Reference
Leukemia	Filter	100(30)	[47]
	Hybrid approach	97.06(3)	[21]
	LOOCV	95.88(3)	
	Hybrid	100(5)	[30]
	PLSVIP	100(9)	[48]
	PLSVEG	100(8)	
	SVM	83.3(3)	[19]
	KNN	97.2(3)	
	NBY	97.2(3)	
	mRMR-ABCs	100(4)	[17]
	GBC	100(8)	
	Clustering	100(10)	[40]
	IDGA-F-SVM	100(15)	[26]
	IDGA-L-NBY	97.7(8.2)	
	IDGA-F-KNN	98.1(13.7)	
	MOBBA-LS		Proposed method
	SVM	97.1(3)	
	KNN	100 (3)	
	NBY	100(3)	
Prostate	MRMR	95.60	[49]
	Embedded	97.0(30)	[50]
	Filter	95.2(30)	[47]
	Hybrid approach	96(8)	[20]
	Clustering	94.71(10)	[40]
	IDGA-F-SVM	96.3(14)	[26]
	IDGA-F-NBY	93.4	
	IDGA-F-KNN	95.6	
	MOBBA-LS		Proposed method
	SVM	94.1(6)	
	KNN	97.1(6)	
	NBY	97.1(6)	
SRBCT	Hybrid approach	100(6)	[21]
	LOOCV	96.04	
	Hybrid	100(8)	[30]
	PLSVIP	100(24)	[48]
	PLSVEG	100(15)	
	SVM	96.4(6)	[19]
	KNN	97.6(6)	
	NBY	97.6(6)	
	mRMR-ABCs	100(6)	[17]
	GBC	95.36(6)	
	IDGA-F-SVM	100(18)	[26]
	IDGA-F-NBY	97.9(29)	
	IDGA-F-KNN	97.8(19)	
	MOBBA-LS		Proposed method
	SVM	85(6)	
	KNN	100(6)	
	NBY	100(6)	

DT stands for Decision Tree classifier, NBY for Naïve Bayes, KNN for K-nearest-neighbor.

Table 6
The resulting p-values of Friedman test.

Methods	NSGA-II	MOBBA-LS
BBA	0.1416	0.0222
NSGA-II	–	0.7300

performance as that of the three-gene subset was observed on training samples, particularly by the SVM. Such observation revealed that if the algorithm was allowed to make more progress; then it perhaps removed the third gene to obtain a smaller set. Thus, the three-gene subset may or not be identified again by the proposed algorithm. That was because in a multi-objective approach, a solution with two genes dominates a solution with three genes when both solutions have an equal performance on the training subset. Overall, it could draw a conclusion that (a): one independent run and applying one classifier did not guarantee the exploration of the most informative subset of genes and (b) obtaining very few number of genes could easily overlook informative genes. On the other hand, achieving perfect prediction accuracy over testing samples and LOOCV does not guarantee a perfect prediction of unseen samples particularly when the number of training samples is scanty. Such issues pose a great challenge for system biologists to gain reliable results perhaps by integrating the machine learning with biological methods. To this sequel, the expression levels of the identified genes are plotted in Fig. 2 which depicts the Cyclin and Macmarcks were highly expressed for ALL while the Zyxin highly expressed for AML.

3.2. Prostate cancer data

The second dataset we used was prostate cancer dataset containing 136 samples consisting of 12600 genes spanning two classes, which includes 59 cancerous tissues and 77 healthy samples [44]. The training set includes 52 prostate cancer samples and 50 normal samples. In addition, the test set contains 25 cancer samples and 9 normal samples. A subset of four highly discriminative genes was obtained by the proposed method. Table 3 presents a description of each gene and the performance of them using four classifiers. The Image-ID 37639_at was already identified in previous studies [40,45,46] as a potential prostate cancer biomarker. The expression levels of identified genes are plotted in Fig. 3, of which only Image-ID 37639_at was highly expressed for normal samples; others were moderately to highly expressed in tumor samples and a few cases of normal ones. From Table 3, the SVM classifier was very well-suited on training samples with zero misclassified sample whereas each of the other three classifiers had only one miss in the normal samples and the other misses in tumor ones. Considering testing samples, KNN, SVM and Naïve Bayes had the same very well performance with only one miss (that was belonged to normal class) out of 34 testing samples. However, DT failed on classifying the testing samples with 25 misses. Overall, the SVM and KNN performed accurately on either the testing samples, training samples or leave-one-out samples.

3.3. SRBCT cancer data

The other dataset we used is the cDNA microarray gene expression profiles of small, round blue cell tumors (SRBCTs) described in [13]. SRBCT dataset includes four different childhood tumors from 83 samples including 25 rhabdomyosarcoma (RMS) samples, 18 neuroblastoma (NB) samples, 29 Ewing's sarcoma (EWS) samples, and 11 Burkitt's lymphoma (BL) samples. From 83 samples, 63 samples (RMS:20, NB:12, EWS:23, BL:8) were used for training. The remaining 20 samples (RMS:20, NB:12, EWS:23, BL:8) were used for blind testing of the system. A subset of six genes was identified by the proposed method from which three genes were already identified in the literature. The Clone id 1435862 was identified as an informative gene in [17] or the 812105 and 244618 were identified in [19]. The expression levels of four highly expressed genes are demonstrated in Fig. 4. Moreover, the performance of the classifiers using the identified subset of genes is reported in Table 4. Using this subset, the SVM and KNN were suited well on training samples with zero misclassified rate. However, on testing samples, the KNN and Naïve Bayes achieved the perfect performance. The SVM had three misclassified samples from three tumor classes (EWS:1, BL:1, RMS:1) in test data. Overall, the KNN performed best using the identified genes even thought that the NBY obtained best LOOCV result with only three misses.

3.4. Comparison with other competitive methods

The performance of the proposed algorithm was compared with some relevant state-of-the-art methods. Table 5 demonstrates the prediction accuracy of the best-identified genes against the best-identified genes of the comparing methods of the literature. Whereas in prostate cancer dataset, the proposed method achieved the highest reported accuracy with a significantly lower number of genes, it significantly

outperformed the literature methods over leukemia dataset with perfect prediction accuracy using only three biomarkers. However, only the method of [17] has the closest performance with ours; with four genes (M31523 at, X62320 at, X66401 cds1 at, M92287 at) and the same accuracy. In SRBCT, the proposed method repeated the perfect results of the literature. We believe that the most remarkable result of our approach concerns the Leukemia dataset where the proposed algorithm satisfied both objectives, i.e. the tiny number of genes and the perfect prediction accuracy. However, the MOBBA-LS attained a comparable performance with the previous studies in other datasets. The time complexity of MOBBA-LS was also studied and compared with the latest binary BA [32] and NSGA-II by allowing each algorithm to progress in five minutes in 10 independent runs over four datasets. The Friedman test statistically corroborated (see Table 6) that MOBBA-LS outperforms BA (under the significance level of 0.05) while having a competitive complexity with NSGA-II that is the most efficient multi-objective algorithm (see Supplementary materials 1 for the results of running times).

4. Concluding marks and future works

The experimental results revealed new combinations of important biomarkers concomitant with the development of three challenging cancers. The explored genes could be further analyzed by investigating their possible role in other diseases. Also, the performance of proposed algorithm could be improved by designing intelligent walk rate and injection probability. The time complexity could also be reduced by manipulating the stopping criterion, for instance, by lowering the number of iterations, and instead, focusing on some ad-hoc analysis upon the final subsets of genes. Furthermore, it would be an invaluable future work to re-implement the algorithm on hardware platform using reconfigurable field programmable gate array (FPGA) such as [54]. Also, it could be implemented on parallel frameworks to dispense its computational burden by getting benefit from distributed computing in cloud environment.

Conflicts of interest

The authors declare no conflict of interests whatsoever.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.ygeno.2017.07.010.

References

[1] J. Ruan, et al., A novel algorithm for network-based prediction of cancer recurrence, Genomics (2016).
[2] S. Fan, K. Huang, R. Ai, M. Wang, W. Wang, Predicting CpG methylation levels by integrating Infinium HumanMethylation450 BeadChip array data, Genomics 107 (4) (2016) 132–137.
[3] M. Mohammadi, H.S. Noghabi, G.A. Hodtani, H.R. Mashhadi, Robust and stable gene selection via maximum–minimum correntropy criterion, Genomics 107 (2) (2016) 83–87.
[4] V. Bhandari, P.C. Boutros, Comparing continuous and discrete analyses of breast cancer survival information, Genomics 108 (2) (2016) 78–83.
[5] C.-Y. Wu, Q.-Z. Li, Z.-X. Feng, Non-coding RNA identification based on topology secondary structure and reading frame in organelle genome level, Genomics 107 (1) (2016) 9–15.
[6] J.M. Moosa, R. Shakur, M. Kaykobad, M.S. Rahman, Gene selection for cancer classification with the help of bees, BMC Med. Genet. 9 (2) (2016) 47.
[7] J. He, M. Sun, Z. Wang, Q. Wang, Q. Li, H. Xie, Characterization and machine learning prediction of allele-specific DNA methylation, Genomics 106 (6) (2015) 331–339.
[8] M. Mohammadi, G.A. Hodtani, M. Yassi, A robust correntropy-based method for analyzing multisample aCGH data, Genomics 106 (5) (2015) 257–264.
[9] A.K. Sharma, A. Gupta, S. Kumar, D.B. Dhakan, V.K. Sharma, Woods: a fast and accurate functional annotator and classifier of genomic and metagenomic sequences, Genomics 106 (1) (2015) 1–6.
[10] P. Guo, et al., Gene expression profile based classification models of psoriasis, Genomics 103 (1) (2014) 48–55.
[11] J. Zahiri, et al., LocFuse: human protein–protein interaction prediction via classifier fusion using protein localization information, Genomics 104 (6) (2014) 496–503.
[12] B. Zhao, B. Xue, Improving prediction accuracy using decision-tree-based meta-strategy and multi-threshold sequential-voting exemplified by miRNA target prediction, Genomics, 2017.
[13] J. Khan, et al., Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks, Nat. Med. 7 (6) (2001) 673–679.
[14] T.R. Golub, et al., Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, Science 286 (5439) (1999) 531–537 (80-.).
[15] Y. Ai-Jun, S. Xin-Yuan, Bayesian variable selection for disease classification using gene expression data, Bioinformatics 26 (2) (2010) 215–222.
[16] V. Bolón-Canedo, N. Sánchez-Maroño, A. Alonso-Betanzos, J.M. Benítez, F. Herrera, A review of microarray datasets and applied feature selection methods, Inf. Sci. 282 (2014) 111–135 (Ny).
[17] H.M. Alshamlan, G.H. Badr, Y.A. Alohali, Genetic bee colony (GBC) algorithm: a new gene selection method for microarray cancer classification, Comput. Biol. Chem. 56 (2015) 49–60.
[18] A.M. Taha, A. Mustapha, S.-D. Chen, Naive bayes-guided bat algorithm for feature selection, Sci. World J. 2013 (2013).
[19] S. Kar, K. Das Sharma, M. Maitra, Gene selection from microarray gene expression data for classification of cancer subgroups employing PSO and adaptive K-nearest neighborhood technique, Expert Syst. Appl. 42 (1) (2015) 612–627.
[20] E.B. Huerta, B. Duval, J.-K. Hao, A hybrid LDA and genetic algorithm for gene selection and classification of microarray data, Neurocomputing 73 (13) (2010) 2375–2383.
[21] Q. Shen, W.-M. Shi, W. Kong, Hybrid particle swarm optimization and tabu search approach for selecting genes for tumor classification using gene expression data, Comput. Biol. Chem. 32 (1) (2008) 53–60.
[22] L. Li, et al., A robust hybrid between genetic algorithm and support vector machine for extracting an optimal feature gene subset, Genomics 85 (1) (2005) 16–23.
[23] I. BoussaïD, J. Lepagnot, P. Siarry, A survey on optimization metaheuristics, Inf. Sci. 237 (2013) 82–117 (Ny).
[24] A. Chakraborty, A.K. Kar, Swarm intelligence: a review of algorithms, Nature-inspired Computing and Optimization, Springer, 2017, pp. 475–494.
[25] X.-S. Yang, A new metaheuristic bat-inspired algorithm, Nature Inspired Cooperative Strategies for Optimization (NICSO 2010), Springer, 2010, pp. 65–74.
[26] M. Dashtban, M. Balafar, Gene selection for microarray cancer classification using a new evolutionary method employing artificial intelligence concepts, Genomics (2017).
[27] J. Xuan, et al., Gene selection for multiclass prediction by weighted fisher criterion, EURASIP J. Bioinforma. Syst. Biol. 2007 (2007) 3.
[28] J. Yang, Y.L. Liu, C.S. Feng, G.Q. Zhu, Applying the Fisher score to identify Alzheimer's disease-related genes, Genet. Mol. Res. GMR 15 (2) (2016).
[29] S. Olyaee, Z. Dashtban, M.H. Dashtban, Design and implementation of super-heterodyne nano-metrology circuits, Front. Optoelectron. 6 (3) (2013) 318–326.
[30] C.-P. Lee, Y. Leu, A novel hybrid feature selection method for microarray data analysis, Appl. Soft Comput. 11 (2011) 208–213.
[31] X.-S. Yang, X. He, Bat algorithm: literature review and applications, Int. J. Bio-inspired Comput. 5 (3) (2013) 141–149.
[32] S. Mirjalili, S.M. Mirjalili, X.-S. Yang, Binary bat algorithm, Neural Comput. Appl. 25 (3–4) (2014) 663–681.
[33] R.Y.M. Nakamura, L.A.M. Pereira, K.A. Costa, D. Rodrigues, J.P. Papa, X.-S. Yang, BBA: a binary bat algorithm for feature selection, Graphics, Patterns and Images (SIBGRAPI), 2012 25th SIBGRAPI Conference on, 2012, pp. 291–297.
[34] S. Mishra, K. Shaw, D. Mishra, A new meta-heuristic bat inspired classification approach for microarray data, Procedia Technol. 4 (2012) 802–806.
[35] K. Deb, A. Pratap, S. Agarwal, T. Meyarivan, A fast and elitist multiobjective genetic algorithm: NSGA-II, Evol. Comput. IEEE Trans. 6 (2) (2002) 182–197.
[36] K.N. Laland, Social learning strategies, Anim. Learn. Behav. 32 (1) (2004) 4–14.
[37] R. Cheng, Y. Jin, A social learning particle swarm optimization algorithm for scalable optimization, Inf. Sci. 291 (2015) 43–60 (Ny).
[38] A. Zhou, B.-Y. Qu, H. Li, S.-Z. Zhao, P.N. Suganthan, Q. Zhang, Multiobjective evolutionary algorithms: a survey of the state of the art, Swarm Evol. Comput. 1 (1) (2011) 32–49.
[39] J. Wang, D. Wang, Particle swarm optimization with a leader and followers, Prog. Nat. Sci. 18 (11) (2008) 1437–1443.
[40] H. Chen, Y. Zhang, I. Gutman, A kernel-based clustering method for gene selection with gene expression data, J. Biomed. Inform. 62 (2016) 12–20.
[41] X. Zhou, D.P. Tuck, MSVM-RFE: extensions of SVM-RFE for multiclass gene selection on DNA microarray data, Bioinformatics 23 (9) (2007) 1106–1114.
[42] K.E. Lee, N. Sha, E.R. Dougherty, M. Vannucci, B.K. Mallick, Gene selection: a Bayesian variable selection approach, Bioinformatics 19 (1) (2003) 90–97.
[43] X. Wang, O. Gotoh, Accurate molecular classification of cancer using simple rules, BMC Med. Genet. 2 (1) (2009) 1.
[44] D. Singh, et al., Gene expression correlates of clinical prostate cancer behavior, Cancer Cell 1 (2) (2002) 203–209.
[45] O. Dagliyan, F. Uney-Yuksektepe, I.H. Kavakli, M. Turkay, Optimization based tumor classification from microarray gene expression data, PLoS One 6 (2) (2011) e14579.
[46] E. Glaab, J. Bacardit, J.M. Garibaldi, N. Krasnogor, Using rule-based machine learning for candidate disease gene prioritization and sample classification of cancer gene expression data, PLoS One 7 (7) (2012) e39932.
[47] L.-J. Zhang, Z.-J. Li, H.-W. Chen, An effective gene selection method based on relevance analysis and discernibility matrix, Pacific-Asia Conference on Knowledge

Discovery and Data Mining, 2007, pp. 1088–1095.
[48] G. Ji, Z. Yang, W. You, PLS-based gene selection and identification of tumor-specific genes, IEEE Trans. Syst. Man Cybern. Part C Appl. Rev. 41 (6) (2011) 830–841.
[49] C. Ding, H. Peng, Minimum redundancy feature selection from microarray gene expression data, J. Bioinform. Comput. Biol. 3 (2) (2005) 185–205.
[50] B. Liu, Q. Cui, T. Jiang, S. Ma, A combinational feature selection and ensemble neural network method for classification of gene expression data, BMC Bioinf. 5 (1) (2004) 1.
[51] Y. Saeys, I. Inza, P. Larrañaga, A review of feature selection techniques in bioinformatics, Bioinformatics 23 (19) (2007) 2507–2517.
[52] V. Piras, K. Selvarajoo, The reduction of gene expression variability from single cells to populations follows simple statistical laws, Genomics 105 (3) (2015) 137–144.
[53] Y. Qi, X. Yang, Interval-valued analysis for discriminative gene selection and tissue sample classification using microarray data, Genomics 101 (1) (2013) 38–48.
[54] J. Kok, L.F. Gonzalez, N. Kelson, FPGA implementation of an evolutionary algorithm for autonomous unmanned aerial vehicle on-board path planning, IEEE Trans. Evol. Comput. 17 (2) (2013) 272–281.