EXCLI Journal 2016;15:460-473 – ISSN 1611-2156
Received: June 27, 2016, accepted: July 12, 2016, published: August 01, 2016

Original article:

CANCER MICROARRAY DATA FEATURE SELECTION USING MULTI-OBJECTIVE BINARY PARTICLE SWARM OPTIMIZATION ALGORITHM

Chandra Sekhara Rao Annavarapu*, Suresh Dara, Haider Banka

Department of Computer Science and Engineering, Indian School of Mines, Dhanbad-826004, Jharkhand, India

* Corresponding author: Chandra Sekhara Rao Annavarapu, Tel: +91-94711 91771, E-mail: acrao401@rediffmail.com, banka.h.cse@ismdhanbad.ac.in

http://dx.doi.org/10.17179/excli2016-481

ABSTRACT

Cancer investigations in microarray data play a major role in cancer analysis and the treatment. Cancer microarray data consists of complex gene expressed patterns of cancer. In this article, a Multi-Objective Binary Particle Swarm Optimization (MOBPSO) algorithm is proposed for analyzing cancer gene expression data. Due to its high dimensionality, a fast heuristic based pre-processing technique is employed to reduce some of the crude domain features from the initial feature set. Since these pre-processed and reduced features are still high dimensional, the proposed MOBPSO algorithm is used for finding further feature subsets. The objective functions are suitably modeled by optimizing two conflicting objectives i.e., cardinality of feature subsets and distinctive capability of those selected subsets. As these two objective functions are conflicting in nature, they are more suitable for multi-objective modeling. The experiments are carried out on benchmark gene expression datasets, i.e., Colon, Lymphoma and Leukaemia available in literature. The performance of the selected feature subsets with their classification accuracy and validated using 10 fold cross validation techniques. A detailed comparative study is also made to show the betterment or competitiveness of the proposed algorithm.

Keywords: Cancer micro array, gene expressions, feature selection, binary PSO, classification

INTRODUCTION

Cancer treatments are targeted for therapies to distinct tumour types by using many computational methods to analyze cancer data, cancer deaths are more than heart disease in persons younger than 85 years (Jemal et al., 2010). Cancer tissue classification is used for diagnosing the cancer. Cancer classification based on gene expression monitoring is used to discover and predict cancer classes of all types without prior biological knowledge (Golub et al., 1999). Prior to classification, finding relevant genes are highly significant to classifying the cancer microarray data. Only few relevant genes are important in the classification. Irrelevant genes cause for low accuracy in classification by hiding relevant features (Guyon et al., 2002). It is therefore not surprising that much effort have been put into developing methods for gene selection (Saeys et al., 2007).

Microarray data, involves the decoding of approximately 30000 human genes, a kind

of NP-Hard problem (Banerjee et al., 2007). Feature selection technique on high dimensional helps to identify key features, also reduces the computational cost and increases the classifier performance. For classifier accuracy in DNA microarray many methods have been proposed, ONCOMINE platform, which is a collection of many gene expression dataset for enlarging its research (Rhodes et al, 2004), recent studies done by (Hatzimichael et al, 2014; Lu et al., 2014) reveals its demand, clustering (Mitra and Ghosh, 2012), and feature selection (Lazar et al., 2012; Linde et al, 2015; Kurakula et al, 2015; Marchan, 2015; Chandrashekar and Sahin, 2014) are recent trends in the research. Hence, expression profiling or microarray gene expression data analyses are prominent tasks in this field.

Feature selection methods selects a subset of 'd' features from a set of 'n' features on the basis of optimization methods. There are many high dimensional datasets which have thousands of features and many of them are irrelevant or redundant. Unnecessary features increase computational burden and make generalization more difficult (Lazar et al., 2012). The feature selection techniques are important tool to reduce dimensionality and to select useful feature subsets that maximizes the classification accuracy (Saeys et al., 2007).

Feature selection methods can be categorized as: filter based, wrapper based, embedded/hybrid based and ensemble methods (Lazar et al., 2012). Filter techniques (Elalami, 2009), selects feature subsets independently of any learning algorithm, assess a significant score with a threshold value to choose the best features. The wrapper model (Sainin and Alfred, 2011) uses predictive accuracy of predetermined learning algorithms. The embedded techniques (Wahid et al, 2011) allow interaction of different class of learning algorithms. More recently, the ensemble model (Nagi and Bhattacharyya, 2013) based on different sub sampling strategies, the learning algorithms run on a number of sub samples and the acquired features are united into a stable subset. However the feature selection techniques can be also categorized based on search strategies used such as forward selection, backward elimination, forward stepwise selection, backward stepwise selection and random mutation (Mladeni, 2006).

Feature selection algorithms are to find feature subsets which are validated by classification accuracy for checking its performance (Yu et al., 2008).

I. Evolutionary computation is a biologically inspired meta-heuristic used for search and optimization representing a powerful and rapidly growing field of artificial intelligence. It uses natural genetics and natural selection to evolve a population of candidate solutions for a given problem. In this paper, we presented a multi objective BPSO algorithm to select feature subsets from high dimensional gene expression data. PSO has certain merits with respect to others such as: i) it uses less number of parameters;
II. it may converge faster and has less computational burden and
III. having potential accuracy.

We proposed a BPSO that preserves better solutions for the next generation. At the first stage, the data is normalized, discretized and converted to binary distinction table by reducing the dimensionality of each sample. At the second stage, BPSO is used to select the significant feature subsets. External validation of selected feature subsets is done in terms of classification accuracy with standard classifiers (Hall et al., 2009).

The remaining part of the paper is structured as follow. The second section describes the preliminaries of PSO algorithm and dominance criteria. The third section discusses about pre-processing of gene expression data, objective functions and the proposed MOBPSO. The fourth section is about various results on three cancerous microarray data such as colon, lymphoma and leukaemia with their validation through standard machine learning classifiers. The last section concludes the article.

EXCLI Journal 2016;15:460-473 – ISSN 1611-2156
Received: June 27, 2016, accepted: July 12, 2016, published: August 01, 2016

PRELIMINARIES

This section formally describes the basics of micro array gene expression data, binary particle swarm optimization algorithm, the dominance criteria with non-dominated sorting algorithm that are relevant for understanding the present work.

Micro array gene expression data

In early 1980's the Array technology was started, did not come into prominence until the mid-1990. But, with the introduction of cDNA microarray technology got lot of fame (Sun et al., 2013). Today, microarrays researchers using array technology in genomic research with a diversified range of applications in biology and medicine. A few recent applications include microbe identification, tumour classification, evaluation of the host cell response to pathogens and analysis of the endocrine system (Konishi et al., 2016).

Analysing DNA microarray data requires a pre-processing phase to produce new biological assumptions, this phase involves distribution, normalization and gene filtering and discretization (Lévêque et al., 2013).

Microarray data classification, which predicts the diagnostic category of a sample from the expression array is a kind of supervised learning. Microarray with orderly arranged samples, provides a good media for matching known and unknown DNA segments with the help of base pairing rules. Microarrays produces huge information requires a series of repeated analyses to render the data interpretable and find out hidden information or pattern in them. The direct output of microarrays is difficult to distinguish various conditions of the samples, or the time points.

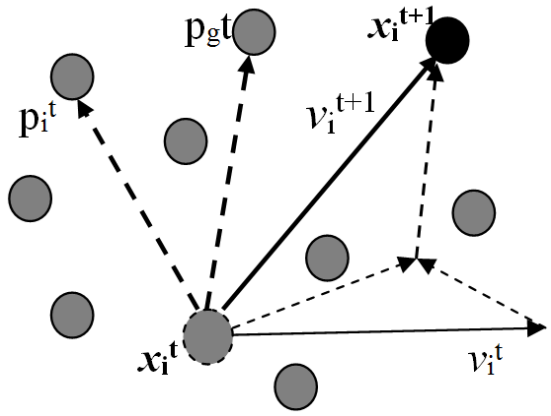

Figure 1: An illustration of PSO architecture

Multi objective optimization

Multi objective optimization involves more than one objective function to get optimal solutions. This involves optimization of single objective function with a trade-off between different objectives, multi objective optimization is also achieved through Particle Swarm Optimization (Coello and Lechuga, 2012).

BPSO (binary particle swarm optimization)

Particle swarm optimization is a heuristic, multi-agent, optimization and evolutionary technique (James and Russell, 1995). It is found to be robust in solving problems featuring nonlinearity, non-differentiability, multi criteria, and high-dimensionality through adaptation which is derived from social-psychological theory (James and Russell, 1997).

The progress of every particle is calculated as per the defined fitness function (James, 1997) and is updated in their velocities and positions according to the following equation.

$$X_{id} = \begin{cases} 1, \text{ if rand}() < S(vid) \\ 0, \text{ otherwise} \end{cases} \quad (1)$$

Where,

$$S(vid) = \frac{1}{(1 + e^{-v_{id}})}$$

Where rand() is a function, to generate a uniform distributed random number in [0,1].

EXCLI Journal 2016;15:460-473 – ISSN 1611-2156
Received: June 27, 2016, accepted: July 12, 2016, published: August 01, 2016

Dominance criteria and non-dominated sorting

In dominance criteria, concept of optimality lies among set of solutions. Solution is said to be dominated with respect to the other solutions based on certain conditions. Non-dominated set in a population are identified with the non-dominated sorting algorithm described in (Deb, 2001).

In this paper, we used two objective functions for finding the non-dominated set among the populations and then assigned the ranks accordingly.

PROPOSED METHODOLOGIES

This section describes the basic preprocessing of gene expression data, objective functions formulation and its justification, followed by the proposed MOBPSO algorithm.

Pre-processing gene expression data

Pre-processing aims to eliminate the ambiguously expressed genes. During feature subset generation, appropriate smallest set of differentially expressed genes are selected across the classes for efficient classification.

1. The normalization is to make the values lie between 0.0 to 1.0. Attribute wise normalization is done by

$$a_j^{|}(x_i) = \frac{(a_i(x_i) - \min_j)}{(\max_j - \min_j)} \quad ,\forall i$$

where max_j maximum and min_j minimum to the gene expression values for attribute a_j over all samples. This makes the normalized continuous attribute value in the range 0 to 1.

2. Then two thresholds Th_i and Th_f, based on the idea of quartiles, are chosen, as in (Banerjee et al., 2007). Let N be the number of patterns in the dataset. The measurements are divided into a number of small class intervals of equal width δ and the corresponding count of class frequencies are fr_c. The position of the k^{th} partition value (k = 1,2,3 for four partitions) is calculated as

$$Th_k = I_c + \frac{(R_k - cfr_{c-1})}{fr_c} * \delta$$

where l_c is the lower limit of the c^{th} class interval, $R_k = \frac{(N * k)}{4}$ is the rank of the k^{th} partition value, and cfr_{c-1} is the cumulative frequency of the preceding class such that $cfr_{c-1} \leq R_k \leq cfr_c$. This has been sketched in the following Figure 2.

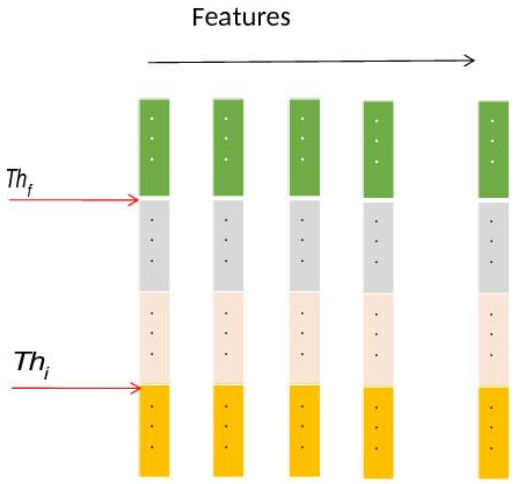

Figure 2: Quartile graph

3. Converting the attribute value table to binary (0/1) form as follows:
4. If $a^{|}(x) \leq Th_i$ Then put '0', Else If $a^{|}(x) \geq Th_f$ Then put '1', Else put '*' (don't care).
5. Find the average occurrences of '*' as threshold Th_a.
6. The attributes whose number of '*'s are $\geq Th_a$ are removed from the table This is the modified (reduced) attribute value table F_r.

After this, the number of features in distinction table becomes 1102 features from 2000 features for colon, become 1867 features from 4026 for lymphoma, and become 3783 features from 7129 for leukemia dataset.

Distinction table preparation

To make a distinction table, a matrix of binary values with dimensions $\frac{(c2 - c1)}{2} * N$

EXCLI Journal 2016;15:460-473 – ISSN 1611-2156
Received: June 27, 2016, accepted: July 12, 2016, published: August 01, 2016

is defined, where N is the number of features in F, C is the number of objects/samples. An entry b((k,j),i) of the matrix corresponds to pair of objects (x_k,x_j) and with the attribute a_i.

$$b((k,j),i) = \begin{cases} 1, \text{ if } a_i(x_k) \neq a_i(x_j). \\ (2) \\ 0, \text{ if } a_i(x_k) = a_i(x_j). \end{cases}$$

The presence of a '1' signifies the attribute a_i's ability to distinguish (or discern) between the pair of objects (x_k, xj).

For a decision table F with N condition attributes and a single decision attribute d, the problem of finding a reduct is equivalent to finding a minimal subset of columns $R(\subseteq \{1,2,\cdots,N\})$ in the distinction table using (4), satisfying $\forall(k,j)\exists i \in R : b((k,j),i) = 1$,whenever $d(x_k) \neq d(x_j)$.

So, in effect, we may consider the distinction table to consist of N columns, and rows corresponding to only those object pairs (x_k,x_j) such that $d(x_k) \neq d(x_j)$.

a) As object pairs corresponding to the same class do not constitute a row of the distinction table, there is a considerable reduction in its size thereby leading to a decrease in computational cost.
b) Additionally, If either of the objects in a pair, has '∗' as an entry under an attribute in table F_r Then in the distinction table, put '0' at the entry for that attribute and pair.
c) The entries '1' in the matrix correspond to the attributes of interest for arriving at a classification decision.

If C_1 and C_2 are the number of objects of the two classes respectively, then rows of the distinction table turn out to be

$$(c_1 * c_2) < \frac{c*(c-1)}{2},$$

where $C_1+C_2 = C$. This reduces time complexity of fitness computation to $O(N * C_1 * C_2)$.

Table 1: A simple example of a distinction table

	F_1	F_2	F_3	F_4	F_5	F_6	F_7
(C_{11},C_{21})	1	1	1	0	1	0	1
(C_{11},C_{22})	0	1	0	1	0	1	0
(C_{11},C_{23})	0	1	1	0	1	0	0
(C_{12},C_{21})	1	0	1	0	1	0	1
(C_{12},C_{22})	0	1	0	0	1	0	0
(C_{12},C_{23})	1	0	1	0	1	0	0

Table 1 describes how a sample distinction table looks like. Here, assume that there is seven conditional features $\{F_1, F_2, F_3, F_4, F_5, F_6, F_7\}$, the length of vector is N = 7. In a vector v, the binary data '1' represents if the corresponding feature is 'present', and a '0' represents its absence. The two classes are C_1 (with two objects i.e. C_{11} and C_{12}) and C_2 (with three objects i.e. C_{21}, C_{22} and C_{23}). The rows represent the object pairs and columns represent the features or attributes. The objective is to choose minimal number of column (features) from the table that covers all the rows (i.e., object pairs in the table). Note that, for multi class problem, if there are k number of classes in a particular dataset, there will be ${}^kC_2$ number of rows in the distinction table. Therefore, the proposed method is not only limited to solve two class problems, but multi-class problem also. However, the present work is focused on two class problems for benchmark datasets as available in literature.

Objective functions design

We used two objective functions Fit_1 and Fit_2. Objective Function 1: The first objective function F_1 is used to finds number of features (i.e. number of 1's). The proposed first objective function is as follow:

$$Fit_1(v) = \frac{(N - O_v)}{N} \quad (3)$$

Objective Function 2: The second objective function F_2 decides the extent to which the feature can recognise among the objects pairs.

EXCLI Journal 2016;15:460-473 – ISSN 1611-2156
Received: June 27, 2016, accepted: July 12, 2016, published: August 01, 2016

$$Fit_2(v) = \frac{R_v}{(c_1 * c_2)} \quad (4)$$

Here, v is the chosen feature subsets, O_v represents the number of 1's in v, C_1 and C_2 are the number of objects in each of the class and R_v is the number of object pairs (i.e. rows in the distinction table) v can discern between. The objective function Fit_1 gives the candidate credit for containing less number of features or attributes in v, and Fit_2 determines the extent to which the candidates can discern among objects pairs in the distinction table.

In simple GA, the two objective functions are combined into one by weighted sum as $Fit = Fit_1 * \alpha + Fit_2 * (1-\alpha)$, where $0 < \alpha_1 < 1$.

As an example to calculate Fit_1 and Fit_2, let us take a sample input vector v = (1,0,1,1,0,1,1), Two classes are C_1 and C_2, where class lengths are $C_1 = 2$, $C_2 = 3$, and length of vector is N = 7 (as depicted in table 1). The number of 1's in v is $O_v = 5$, and R_v is calculated as compare with input vector v matching number of presented 1's from each row in distinction table, i.e $R_v = 5$. Therefore

$$Fit_1(v) = \frac{(N - O_v)}{N} = (7-5)/\,7 = 0.29,$$

and

$$Fit_2(v) = \frac{R_v}{(c_1 * c_2)} = 5\,/6 = 0.84.$$

Here, a multi objective BPSO algorithm is proposed for feature subset selection. The best non-dominated solutions of combined population of swarm at two successive generations (i.e., current and next population) are preserved at each generation. Only best 50 % solutions are allowed to evolve for the next generation. This is repeated for finite number of generations. The proposed approach is described in Algorithm [1].

The MOBPSO algorithm for feature selection

Algorithm 1: The proposed Multi Objective BPSO algorithm

Step: 1 Initialize P no. of solutions with random velocity and positions
Step: 2 Calculate fitness on P using equation (5) and (6)
Step: 3 Update pbest Update gbest
Step: 4 Update velocities and coordinates of P using equation (1) and (3) to generate P^{I}
Step: 5 Calculate fitness values for P^{I} using equation (5) and (6)
Step: 6 Combine both P and P^{I} as P^{II}
Step :7 Perform non-dominated sorting on P^{II} using algorithm [1]
Step: 8 Choose 50 % best ranked solutions from P^{II} as P
Step: 9 Repeat Step (3–8) for finite number of generations

gbest selection

After perform non-dominating sorting on mixed population (i.e. parent and child), we can get non dominated solutions. Here, we choose one random solution as gbest among top raked non-dominated solutions. Since, more than one top ranked solutions are may be available, but all solutions having same priority.

RESULTS AND DISCUSSIONS

Cancer gene expression data sets

In this study, three benchmark cancer datasets have been used for training and testing purpose.

I. Colon Cancer dataset available at (http://genomics-pubs.princeton.edu/oncology/) is a set of 62 gene expressions, containing 2000 genes (features).
II. Lymphoma dataset available at (http://llmpp.nih.gov/lymphoma/data/figure1/figure1.cdt) is a set of 96 gene expressions, having 4026 genes.
III. Leukaemia dataset available at (http://www.genome.wi.mit.edu/MPR) is a set of 38 gene expressions, having 7129 genes.

EXCLI Journal 2016;15:460-473 – ISSN 1611-2156
Received: June 27, 2016, accepted: July 12, 2016, published: August 01, 2016

RESULTS

Our proposed MBPSO on colon microarray, lymphoma microarray, and leukemia microarray obtained minimal subsets of features. In our experiment values of accelerator coefficients c_1 and c_2 are set to 2 whereas velocities set to minimum of -4 and maximum of 4 (Sudholt and Witt, 2008). In BPSO, inertia weight (w) treated as one of the most important parameter, through which we can improve accuracy by estimation and balancing of local and global search (Shi and Eberhart, 1999). After several experiments, 'w' was set to 0.9. Various population sizes were taken, to check feature subsets behaviour, also the swarm size set as per literature. After several experiments maximum number of runs was set to 50 which were also tested with varied population size like 10, 20, 30 and 50. Many standard classifiers have been used for testing purposes to show consistent performance and robustness of the proposed method. The experimental results are carried out on three bench mark datasets as summarized in Table 3.

Note that k is chosen to be an odd number to avoid the ties. The correct classification are reported to be 93.54 %, 95.85 % and 94.74 % for those three datasets with varies swarm size and k values. The results shown above is based on average score over (10-15) runs. Table 3 represents k−NN classification results with single objective function of GA. Here, it is giving 100 % correct classification score for all three data sets when k = 1. For colon data 93.55 % score when k = 3, 90.33 % for k = 5 and for k = 7 it is 83.88 %, on 10 feature subset.

Table 2: Details of the cancer microarray datasets before and after pre-processing

Datasets	Total Features	Reduced Features#	Classes	Samples
Colon	2000	1102	Colon	40
			Normal	22
Lymphoma	4026	1867	Other	54
			B-cell	42
Leukemia	7129	3783	ALL	47
#-after preprocessing			AML	25

Table 3: K-nearest neighbour (k-NN) classification results on colon microarray, lymphoma microarray, and leukemia microarray for the performance with proposed method

Dataset	Population Size	Subset features	k-NN) classification (%) on test set			
			K = 1	K = 3	K = 5	K = 7
Colon: #Genes 2000 Reduce to 1002	10	10	100	83.87	83.87	80.65
	20	9	100	83.87	83.87	83.87
	30	9	100	93.54	80.65	83.87
	50	9	100	90.32	80.65	87.09
Lymphoma #Genes 4026 Reduce to 1867	10	20	100	93.75	93.75	89.75
	20	22	100	95.85	91.66	91.66
	30	21	100	95.85	93.75	91.66
	50	15	100	93.75	93.75	91.66
Leukemia #Genes 7129 Reduce to 3783	10	14	100	89.49	89.49	89.49
	20	15	100	92.10	89.49	92.10
	30	14	100	94.75	89.49	89.49
	50	14	100	94.75	86.85	89.49

EXCLI Journal 2016;15:460-473 – ISSN 1611-2156
Received: June 27, 2016, accepted: July 12, 2016, published: August 01, 2016

For lymphoma data, it is 93.75 %, 93.75 % and 89.59 % where k = 3, k = 5 and k = 7 respectively. Similarly, for leukemia data, where k = 3, 5 and 7 the correct classification is 94.74 %.

Table 4 depicts the results for the all three datasets using Bayes Family classifiers. The Bayes Logistic Regression (BLR), Bayes Net(BN) classifier, and Naive Bayes(NB) classifiers given 93.55 % highest correct classification result on 13 features subset for colon data. For the lymphoma data, 100 % correct classification score has been achieved by using Bayes Logistic Regression whereas Bayes Net classifier gives 95.84 % and Naive Bayes classifiers gives 97.42 % on 22 feature subset. Similarly for leukemia data, it is 92.1 % classification with Bayes Logistic Regression, 89.48 % with other two classifiers on 14 feature subset.

We investigated different well known function based classifiers such as LibLinear, LibSVM, Logistic, Multilayer perceptron (MLP), stochastic gradient descent (SGD) and Spegasos, reported in Table 5. For colon data, 100 % correct classification score using all classifiers with 13 and 9 gene subsets except SGD and SPegasos classifiers, where those are giving 96.78 % as highest classification on 13 gene subset. For lymphoma data, same table depicts, 100 % score for all classifiers with various (i.e. 15 to 22) gene subset. For leukemia data, shows that 100 % correct classification score for all classifiers with various gene subsets except SGD and SPegasos classifiers, where those are giving 94.71 % and 93.37 % as highest classification on 13 gene subset.

Table 6 shows that results of various well known Tree based classifiers such as Best First Decision (BFT), Decision Tree (DT), Functional tree (FT), Decision tree classifier (J48), Logistic model tree (LMT), Randomforest (RF) and reduced error pruning tree (REPT). From the table we observe that, except BFT and DT classifiers, remaining all classifiers are giving 100 % correct classification score at various selected subsets for all three data sets. The BFT giving 93.55 % correct score for colon data, 97.91 % for lymphoma data, and 97.37 % for leukemia data. And the Decision Stump classifiers are giving 83.88 % for colon data with 10 and 6 gene subset, 87.50 % for lymphoma data with 14 gene subset and it is 86.85 % for leukemia data with 11 gene subset. We achieved 100 % correct classification on some not included in the Table 6 Alternating Decision Tree, Extra Tree, LADTree and Random Tree classifiers. The classifiers are shown in Table 6.

Table 4: Performance on three datasets using Bayes family Classifiers

Dataset	Selected features	Used classifier Method		
		BLR	Bayes Net	Naive Bayes
Colon	13	93.55	93.55	93.55
	9	87.10	67.74	80.65
	9	87.1	67.74	80.64
	9	80.65	64.52	83.88
Lymphoma	20	95.84	93.75	93.75
	22	100	95.84	97.42
	21	95.84	91.67	93.75
	15	93.75	89.59	93.75
Luekemia	14	92.2	71.1	89.48
	15	86.85	86.85	89.48
	14	92.1	89.48	89.48
	14	89.48	89.48	89.48

EXCLI Journal 2016;15:460-473 – ISSN 1611-2156
Received: June 27, 2016, accepted: July 12, 2016, published: August 01, 2016

Table 5: Performance on three datasets using Function Based Classifiers

Dataset	Selected features	Used classifier methods and results in (%)					
		LibLinear	LibSVM	Logistic	MLP	SGD	Spegasos
Colon	13	100	100	96.78	100	96.78	96.78
	9	93.54	100	100	96.77	90.32	93.55
	9	93.54	100	100	96.77	90.32	93.55
	9	77.42	100	90.33	100	83.87	90.32
Lymphoma	20	100	93.75	100	97.92	97.92	100
	22	100	100	100	97.92	100	100
	21	100	95.84	100	97.92	95.84	100
	15	100	91.67	100	100	95.84	100
Leukemia	14	100	100	100	97.37	94.74	86.84
	15	92.11	100	100	97.37	94.74	92.11
	14	89.48	100	100	100	92.1	97.37
	14	100	100	100	100	89.48	97.37

Table 6: Performance on three datasets using Tree Based Classifiers

Dataset	Selected features	Used classifier methods and result in (%)						
		BFT	DS	FT	J48	LMT	RF	REPT
Colon	13	96.78	87.10	100	96.78	100	100	96.78
	9	64.52	67.75	87.1	96.78	87.1	100	96.78
	9	64.52	67.75	87.1	96.78	87.1	100	96.78
	9	93.55	77.42	87.09	96.77	83.87	100	77.42
Lymphoma	20	93.75	87.5	93.75	97.91	93.55	97.91	93.75
	22	93.75	87.5	100	97.92	100	100	87.5
	21	95.84	87.5	97.92	95.84	91.67	100	91.67
	15	91.67	81.25	97.92	95.84	91.67	91.67	81.25
Leukemia	14	89.49	71.1	94.74	97.37	97.37	100	86.84
	15	92.11	84.22	94.74	92.11	92.11	100	92.11
	14	97.37	89.48	92.1	97.37	86.85	100	89.48
	14	97.37	71.1	100	97.37	97.37	100	71.1

K-fold cross validation

Cross-validation techniques are a thorough computational mechanism to estimate performance by the use of examples as training and testing sets. In K-fold cross validation, mimics the training and test sets by repeatedly training the algorithm K times with a fraction 1/K of training examples left out for testing purposes. We use K = 10, which is also called 10-fold cross validation, in each experimental run, nine folds are used for training and remaining one fold is used for testing. Therefore, training and test sets consist of 90 % and 10 % of data (Zhang, 2011). Our 10-fold cross validation result on

EXCLI Journal 2016;15:460-473 – ISSN 1611-2156
Received: June 27, 2016, accepted: July 12, 2016, published: August 01, 2016

colon, lymphoma and leukemia datasets reported in Table 7.

Table 7: 10-fold cross validation on colon, lymphoma and leukemia

Dataset	Classification		Standard	
	correct	miss	mean	deviation
Colon	84.52	15.48	0.155	1.128
Lymphoma	91.67	8.33	0.084	0.851
Leukemia	80.69	19.31	0.194	1.025

Comparisons

(Liang et al., 2013) introduces a new evolving personalized modelling method and system (evoPM) that integrates gravitational search inspired algorithm (GSA) for selecting informative features. Here, they used 4 high dimensional benchmark datasets, and reported the selected feature subsets with a minimum of 25 features and maximum of 101 features. In our algorithm, selected feature subsets are 9 to 22. Moreover, their classification accuracy is 87.1 % for colon, 94.81 % for lymphoma, and it was 97.22 % for leukemia data. The proposed MOBPSO performs 100 % classification by some of the classifiers on these datasets.

In Figure 3, Performance of Proposed MOBPSO Algorithm, NSGA-II and GA on colon, lymphoma and leukemia datasets respectively using bayes Classifiers such as BLR, BN and NB are shown. In Figure 4, Performance of Proposed MOBPSO Algorithm, NSGA-II and GA on colon, lymphoma and leukemia datasets respectively using Function based Classifiers such as LibLinear, LibSVM, Logistic, MLP, SGD and SPegasos are shown. In Figure 5, Performance of Proposed MOBPSO Algorithm, NSGA-II and GA on colon, lymphoma and leukemia datasets using Tree based Classifiers such as BFT, DS, FT, J48, LMT, RF and REPT are shown. Figure 6 demonstrates of heat maps for three datasets with reduced feature subsets of gene samples. The heat map is graphical representation of data to represent the level of expression of many genes across a number of comparable samples as they are obtained from DNA microarrays, where the individual values contained in a matrix are represented as colours. Larger values were represented by small dark gray or black colour and smaller values by lighter colours.

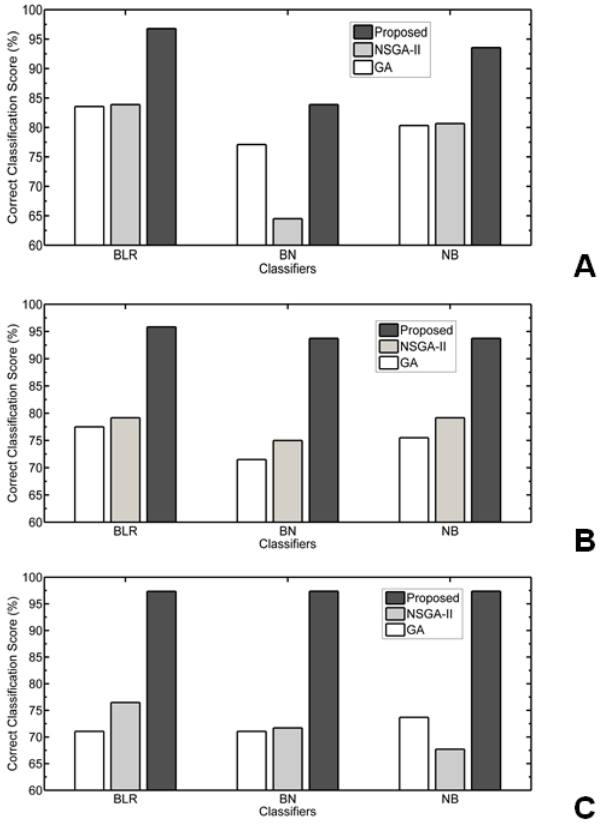

Figure 3: Performance of Proposed MOBPSO Algorithm, NSGA-II and GA on three datasets using bayes Classifiers (BLR, BN, NB); **(A)** performance of Proposed MOBPSO Algorithm, NSGA-II and GA on Colon dataset using bayes Classifiers (BLR, BN, NB), **(B)** performance of Proposed MOBPSO Algorithm, NSGA-II and GA on lymphoma dataset using bayes Classifiers (BLR, BN, NB), **(C)** performance of Proposed MOBPSO Algorithm, NSGA-II and GA on leukemia dataset using bayes Classifiers (BLR, BN, NB)

EXCLI Journal 2016;15:460-473 – ISSN 1611-2156
Received: June 27, 2016, accepted: July 12, 2016, published: August 01, 2016

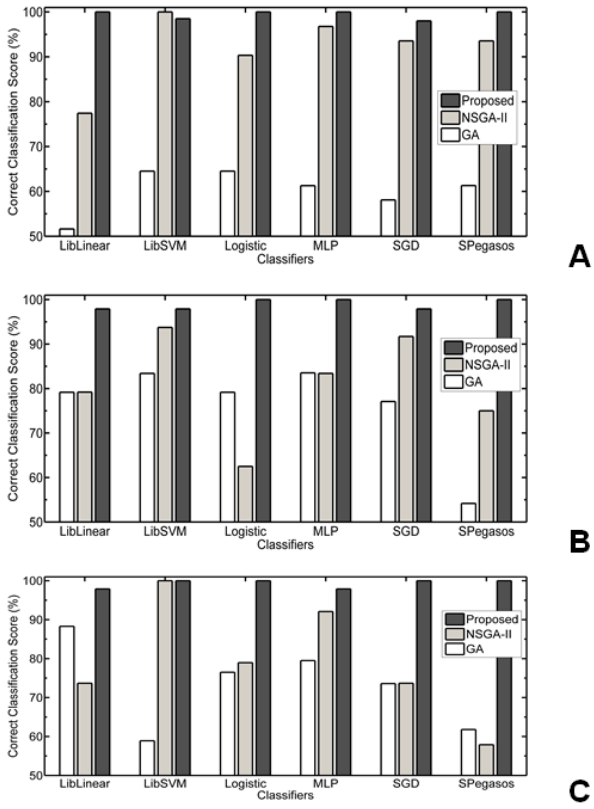

Figure 4: Performance of Proposed MOBPSO Algorithm, NSGA-II and GA on three datasets using Function based Classifiers; **(A)** performance of Proposed MOBPSO Algorithm, NSGA-II and GA on Colon dataset using Function based Classifiers**, (B)** performance of Proposed MOBPSO Algorithm, NSGA-II and GA on lymphoma dataset using Function based Classifiers**, (C)** performance of Proposed MOBPSO Algorithm, NSGA-II and GA on leukemia dataset using Function based Classifiers

Z-score analysis

Z scores provide a relative, semi quantitative estimation of microarray gene expression levels. Z score is calculated on the basis of hybridized intensity among experiments of same array type. Z score reflections on different hybridization values are as follows:

I. Z scores values with higher positive represent the genes with high expressiveness
II. Z scores values with Low negative values represent genes that are least expressed (Cheadle et al., 2003).

Z scores are mathematically calculated as follows:

$$Z = \frac{(G_i - \mu)}{\sigma} \quad (7)$$

where G_i intensity of Gene, μ is mean of intensity $G_1 ... G_n$ (i.e. aggregate measure of all genes), and $\sigma = \sum_{i=1}^{n} G_i$ is SD.

In Figure 7, Z score of colon, lymphoma and leukemia datasets, with a selected feature subsets and the selected genes, which are highly expressed to select are shown.

CONCLUSION AND FUTURE SCOPE

In this paper, we presented a MOBPSO to find feature subset in cancer gene expression microarray data sets. Non-dominating sorting helps to preserve Pareto-font solutions. The pre-processing aids faster conver-

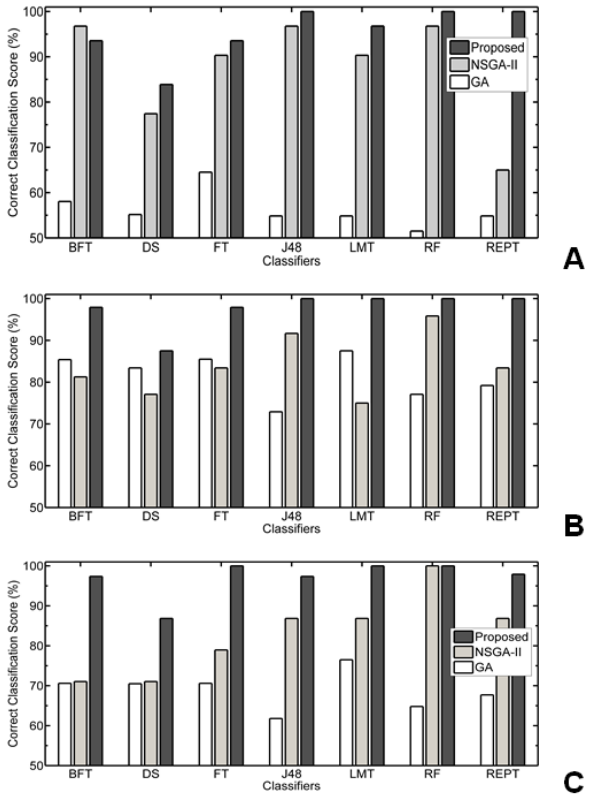

Figure 5: Performance of Proposed MOBPSO Algorithm, NSGA-II and GA on three datasets using Tree based Classifiers; **(A)** performance of Proposed MOBPSO Algorithm, NSGA-II and GA on Colon dataset using Tree based Classifiers, **(B)** performance of Proposed MOBPSO Algorithm, NSGA-II and GA on lymphoma dataset using Tree based Classifiers, **(C)** performance of Proposed MOBPSO Algorithm, NSGA-II and GA on leukemia dataset using Tree based Classifiers

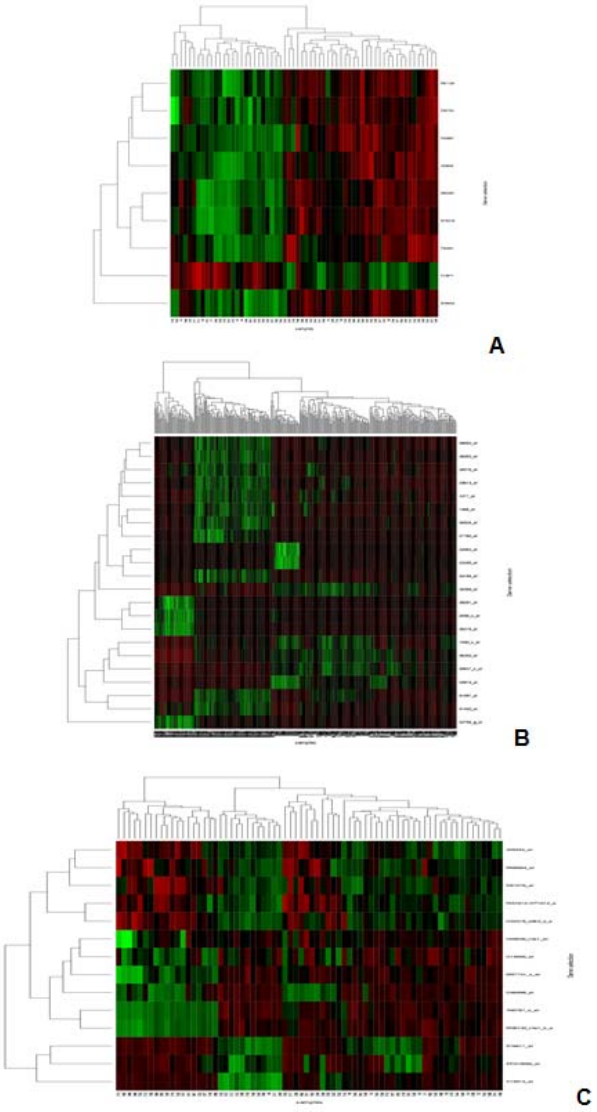

Figure 6: Heat map of three datasets with reduced features; **(A)** heat map on colon data with reduced features having 9 genes, **(B)** heat map on lymphoma data with reduced features having 22 genes, **(C)** heat map of Leukemia data with reduced features having 14 genes

gence along the search space and successfully employed to eliminate redundant and irrelevant features. The proposed approach is experimentally investigated with different parameters. The main goal of the feature selection is selecting minimal feature subsets with higher classification accuracy which has been achieved by two objective functions. The result on three benchmark cancer datasets demonstrates the feasibility and effectiveness of the proposed method. The performances of the proposed along with the existing methods are compared using standard classifiers and reported better and competitive performance.

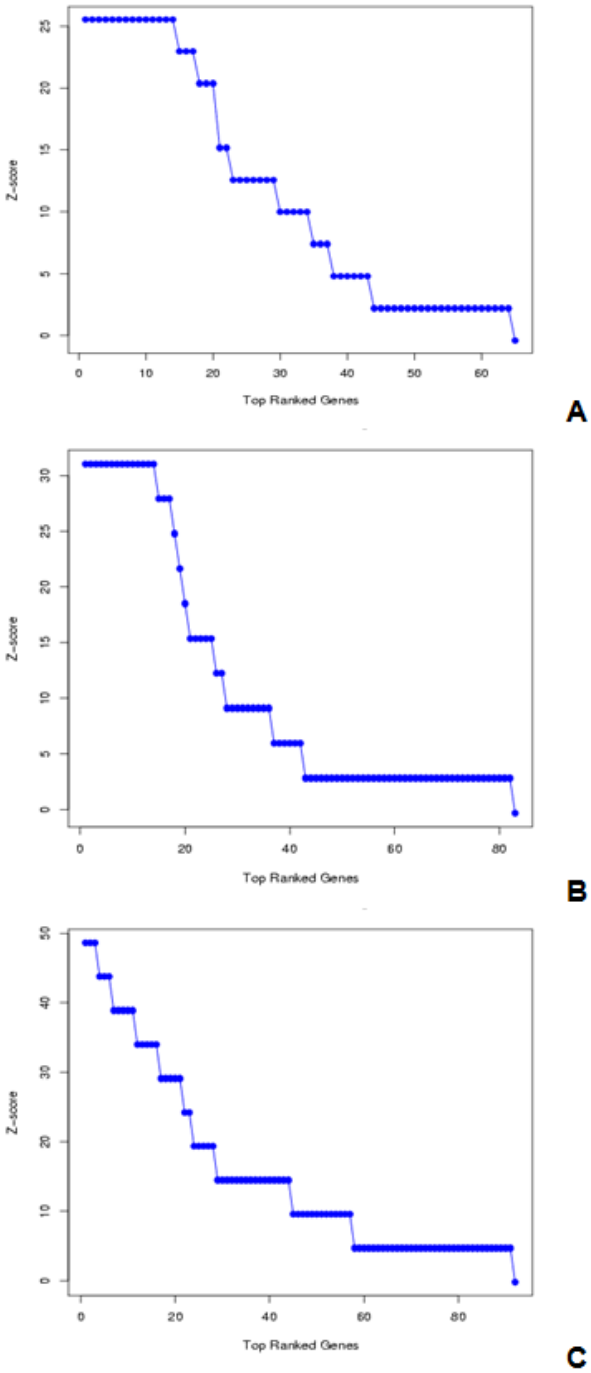

Figure 7: Z-Score Analysis of colon, lymphoma and leukemia having 9, 22, and 14 genes respectively; **(A)** Z-Score Analysis of colon having 9 genes, **(B)** Z-Score Analysis of Lymphoma having 22 genes, **(C)** Z-Score Analysis of Leukemia having 14 genes

Conflict of interest

None declared.

REFERENCES

Banerjee M, Mitra S, Banka H. Evolutionary rough feature selection in gene expression data. IEEE Trans Syst Man Cybern Part C (Applications and Reviews). 2007;37:622-32.

Chandrashekar G, Sahin F. A survey on feature selection methods. Comput Elec Engin 2014;40:16-28.

Cheadle C, Vawter MP, Freed WJ, Becker KG. Analysis of microarray data using Z score transformation. J Mol Diagnost. 2003;5:73-81.

Coello CC, Lechuga MS. MOPSO: A proposal for multiple objective particle swarm optimization. In: Proceedings of the 2002 Congress on Evolutionary Computation, 12-17 May 2002, Honolulu, HI/USA. CEC'02, Vol. 2 (pp 1051-6). NewYork: IEEE, 2002.

Deb K. Multi-objective optimization using evolutionary algorithms. Chichester: Wiley, 2001.

Elalami ME. A filter model for feature subset selection based on genetic algorithm. Knowledge-Based Systems. 2009;22:356-62.

Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science. 1999;286:531-7.

Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Machine Learning. 2002;46:389-422.

Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: an update. ACM SIGKDD Explorations Newsletter 2009;11(1):10-8.

Hatzimichael E, Lagos K, Van Ren Sim EB, Crook T. Epigenetics in diagnosis, prognostic assessment and treatment of cancer: an update. EXCLI J. 2014;13: 954-76.

James K. The particle swarm: social adaptation of knowledge. In: Proceedings of 1997 IEEE International Conference on Evolutionary Computation (ICEC '97), 13-16 April 1997, Indianapolis, IN/USA (pp 303-8). Piscataway, NJ: IEEE, 1997.

James K, Russell E. Particle swarm optimization. In: Proceedings of IEEE International Conference on Neural Networks, 27.11.-1.12.1995 (pp 1942-8). Piscataway, NJ: IEEE, 1995.

James K, Russell E. A discrete binary version of the particle swarm optimization. In: Proceedings of the Conference on System, Man, and Cybernetics (pp 4104-8). Piscataway, NJ: IEEE, 1997.

Jemal A, Siegel R, Xu J, Ward E. Cancer statistics, 2010. CA: a cancer journal for clinicians. 2010;60: 277-300.

Konishi H, Ichikawa D, Arita T, Otsuji E. Microarray technology and its applications for detecting plasma microrna biomarkers in digestive tract cancer. Meth Mol Biol. 2016;1368:99-109.

Kurakula K, Goumans MJ, ten Dijke P. Regulatory RNAs controlling vascular (dys) function by affecting TGF-ß family signalling. EXCLI J. 2015;14:832-50.

Lazar C, Taminau J, Meganck S, Steenhoff D, Coletta A, Molter C, et al. A survey on filter techniques for feature selection in gene expression microarray analysis. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB). 2012;9(4):1106-19.

Lévêque N, Renois F, Andréoletti L. The microarray technology: facts and controversies. Clin Microbiol Infect. 2013;19:10-4.

Liang W, Hu Y, Kasabov N. Evolving personalized modeling system for integrated feature, neighborhood and parameter optimization utilizing gravitational search algorithm. Evolv Syst. 2015;6:1-4.

Linde J, Schulze S, Henkel SG, Guthke R. Data- and knowledge-based modeling of gene regulatory networks: an update. EXCLI J. 2015;14:346-78.

Lu HJ, An CL, Zheng EH, Lu Y. Dissimilarity based ensemble of extreme learning machine for gene expression data classification. Neurocomputing. 2014; 128:22-30.

Marchan R. Highlight report: Validation of prognostic genes in lung cancer. EXCLI J. 2014;13:457-60.

Mladenic D. Feature selection for dimensionality reduction. In: SLSFS'05. Proceedings of the 2005 international conference on Subspace, Latent Structure and Feature Selection (pp 84-102). Berlin: Springer-Verlag, 2006 (Lecture notes in computer science, Vol. 3940).

Mitra S, Ghosh S. Feature selection and clustering of gene expression profiles using biological knowledge. IEEE Trans Syst Man Cybern, Part C (Applications and Reviews). 2012;42:1590-9.

Nagi S, Bhattacharyya DK. Classification of microarray cancer data using ensemble approach. Network Model Anal Health Inform Bioinform. 2013;2:159-73.

Rhodes DR, Yu J, Shanker K, Deshpande N, Varambally R, Ghosh D et al. ONCOMINE: a cancer microarray database and integrated data-mining platform. Neoplasia. 2004;6(1):1-6.

Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. Bioinformatics. 2007;23:2507-17.

Sainin MS, Alfred R. A genetic based wrapper feature selection approach using nearest neighbour distance matrix. In: 2011 3rd Conference on Data Mining and Optimization, Putrajaya, Malaysia, June 2011 (pp 237-42). Piscataway, NJ: IEEE, 2011.

Shi Y, Eberhart RC. Empirical study of particle swarm optimization. In: Congress on Evolutionary Computation, CEC 99, July 6-9, 1999, Washington, DC (pp 1945-50). Piscataway, NJ: IEEE, 1999.

Sudholt D, Witt C. Runtime analysis of binary PSO. In: GECCO '08. Genetic and Evolutionary Computation Conference, Atlanta, GA/USA, July 12-16, 2008 (pp 135-42). New York: ACM, 2008.

Sun H, Chen GY, Yao SQ. Recent advances in microarray technologies for proteomics. Chem Biol. 2013; 20:685-99.

Wahid CM, Ali AS, Tickle KS. A novel hybrid approach of feature selection through feature clustering using microarray gene expression data. In: Proceedings of the 2011 11th International Conference on Hybrid Intelligent Systems (HIS), 5.-8.12.2011, Malacca, Malaysia (pp 121-6). Piscataway, NJ: IEEE, 2011.

Yu L, Ding C, Loscalzo S. Stable feature selection via dense feature groups. In: KDD '08. The 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, NV/USA, Aug. 24-27, 2008 (pp 803-11). New York: ACM, 2008.

Zhang F. Cross-validation and regression analysis in high-dimensional sparse linear models. Thesis. Stanford, CA: Stanford Univ., 2011.