Expert Systems with Applications 42 (2015) 612–627

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Gene selection from microarray gene expression data for classification of cancer subgroups employing PSO and adaptive *K*-nearest neighborhood technique

Subhajit Kar [a,*], Kaushik Das Sharma [b], Madhubanti Maitra [c]

[a] Department of Electrical Engineering, Future Institute of Engineering and Management, Kolkata, India
[b] Department of Applied Physics, University of Calcutta, Kolkata, India
[c] Department of Electrical Engineering, Jadavpur University, Kolkata, India

ARTICLE INFO

Article history:
Available online 19 August 2014

Keywords:
Microarray data
SRBCT data
ALL_AML data
MLL data
Particle swarm optimization (PSO)
Adaptive *K*-nearest neighborhood (KNN)
Support vector machine (SVM)

ABSTRACT

These days, microarray gene expression data are playing an essential role in cancer classifications. However, due to the availability of small number of effective samples compared to the large number of genes in microarray data, many computational methods have failed to identify a small subset of important genes. Therefore, it is a challenging task to identify small number of disease-specific significant genes related for precise diagnosis of cancer sub classes. In this paper, particle swarm optimization (PSO) method along with adaptive *K*-nearest neighborhood (KNN) based gene selection technique are proposed to distinguish a small subset of useful genes that are sufficient for the desired classification purpose. A proper value of *K* would help to form the appropriate numbers of neighborhood to be explored and hence to classify the dataset accurately. Thus, a heuristic for selecting the optimal values of *K* efficiently, guided by the classification accuracy is also proposed. This proposed technique of finding minimum possible meaningful set of genes is applied on three benchmark microarray datasets, namely the small round blue cell tumor (SRBCT) data, the acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML) data and the mixed-lineage leukemia (MLL) data. Results demonstrate the usefulness of the proposed method in terms of classification accuracy on blind test samples, number of informative genes and computing time. Further, the usefulness and universal characteristics of the identified genes are reconfirmed by using different classifiers, such as support vector machine (SVM).

1. Introduction

Gene expression data are ever increasingly being used in the field of medicine including categorization of carcinoma into different histopathological subgroups, which often appear almost identical in orthodox histopathology microscopically done in pathological laboratories (Pal, Aguan, Sharma, & Amari, 2007). In this regard, computer-aided image analysis could help in corroborating the findings of the histopathologists. Computer aided image analysis for cancer classification actually rests on the efficacious analysis of gene expression data. However, analysis of gene expression data poses a major problem due to the presence of huge number of genes in microarray dataset. The analysis becomes even more difficult as the numbers of accessible samples available are alarmingly less. Therefore, identification of a small subset of informative genes from the large number of gene-set for classifying the samples accurately turns out to be a demanding job. To overcome this challenge, gene selection methods are usually divided into two categories (Mohamad, Omatu, Deris, & Yoshioka, 2011). The first one that is the filter methods perform gene selection in a-priori basis before the dataset is used for classification analysis. In contrast, the wrapper methods search for the best genes in the space of all gene subsets at the time of classification. Filter methods (Chen, Liu, Ma, & Hua, 2005; Furey et al., 2000; Xiong, Fang, & Zhao 2001; Xiong, Li, Zhao, Li, & Boerwinkle 2001) are usually known as gene-ranking methods, which include t-test (Shen, Shi, & Kong, 2008), gain ratio (Mohamad et al., 2011), Wilcoxon rank sum test (Li, Wu, & Tan, 2008) and these methods are computationally more efficient than the wrapper methods (Xiong et al., 2001; Xiong & Li et al., 2001). However, by using gene-ranking methods, some genes among the selected genes may come out to be redundant because they contribute no additional information towards the subset. They have similar expression levels among

* Corresponding author. Tel.: +91 9002299226.
E-mail address: sksubhajit@gmail.com (S. Kar).

http://dx.doi.org/10.1016/j.eswa.2014.08.014
0957-4174/

Table 1
Experimental result for each run on SRBCT dataset.

Run	No. of Genes	Index no. of selected genes	Cross validation Acc.	*K* for KNN	Training Acc.	Test Acc.	Computing Time (Hrs.)
1	13	187,359,390,742,1051,1074,1374,1738,1841,1842, 1947, 1956,1980	98.0159	6	98.4127 (62/63)	90 (18/20)	5.5859
2	13	107,123,246,518,714,742,1181,1206,1286,1343,1954,2097,2106	99.2063	4	100 (63/63)	90 (18/20)	2.8632
3	5	246,257,742,1076,2103,	93.2540	4	93.6508 (59/63)	90 (18/20)	2.8098
4	10	187,231,246,356,428,534,650,742,758,840	95.6349	6	95.2381 (60/63)	90 (18/20)	2.7993
5	6	742,1003,1386,2046,2099,2157	98.0159	4	100 (63/63)	100 (20/20)	2.7956
6	9	125,187,246,362,742,783,1496, 1764,2040	96.8254	3	96.8254 (61/63)	90 (18/20)	2.8111
7	7	742,1601,1645,1932,1955,2046,2144	93.2540	3	100 (63/63)	100 (20/20)	2.8202
8	7	84,251,742,1032,1074,1645,1783	94.4444	4	100 (63/63)	100 (20/20)	2.8098
9	9	255,694,806,1003,1117,1319,1477,2046,2113	96.8254	3	100 (63/63)	95 (19/20)	2.8017
10	6	246,841,1263,1531,1955, 1645	94.8413	3	96.8254 (61/63)	95 (19/20)	2.7981
Avg.	8.5		96.0318		98.0952	94	3.0895
Std. Dev.	2.8382		2.0746		2.3424	4.5947	0.8774

Table 2
Experimental result for each run on ALL_AML dataset.

Run	No. of Genes	Index no. of selected genes	Cross Validation Acc.	*K* for KNN	Training Acc.	Test Acc.	Computing Time (Hrs.)
1	3	804,1249,1334	97.4206	4	100 (38/38)	94.1176 (32/34)	2.8065
2	3	804,1882,4052	95.8868	4	100 (38/38)	97.0588 (33/34)	2.7906
3	2	1882,4108	95.8333	3	100 (38/38)	94.1176 (32/34)	2.4086
4	4	804,1882,4052,2642	100	4	100 (38/38)	97.0588 (33/34)	2.8065
5	4	187,1096,1882,3747	95.1389	3	97.3684 (37/38)	94.1176 (32/34)	2.8065
6	2	2642,4052	97.9701	6	97.3684 (37/38)	76.4706 (26/34)	2.7973
7	3	804,1882,2642	98.6111	4	100 (38/38)	97.0588 (33/34)	2.7956
8	2	1882,4052	94.5513	6	94.7368 (36/38)	94.1176 (32/34)	2.6201
9	2	1882,2642	97.2222	4	100 (38/38)	94.1176 (32/34)	2.4086
10	2	804,1882	98.6111	4	97.3684 (37/38)	97.0588 (33/34)	2.7906
Avg.	2.7		97.1245		98.6842	93.5294	2.7031
Std. Dev.	0.8233		1.7405		1.8608	6.1695	0.1651

Table 3
Experimental result for each run on MLL dataset.

Run	No. of Genes	Index no. of selected genes	Cross validation Acc.	*K* for KNN	Training Acc.	Test Acc.	Computing Time (Hrs.)
1	4	2835,4341,8165,8937	98.6842	6	100 (57/57)	93.3333 (14/15)	7.1088
2	3	4341,2028,12418	86.4035	6	94.7368 (54/57)	93.3333 (14/15)	7.2089
3	4	3479,657,841,7930	76.3158	3	91.2281 (52/57)	86.6667 (13/15)	7.2917
4	4	841,7930,12418,4341	88.1579	4	96.4912 (55/57)	93.3333 (14/15)	7.3956
5	4	841,7930,8937,12418	92.5439	4	100 (57/57)	100 (15/15)	7.1488
6	4	4341,7930,2028,841	87.7193	6	91.2281 (52/57)	86.6667 (13/15)	7.1667
7	3	657,7136,12418	85.0877	4	96.4912 (55/57)	86.6667 (13/15)	7.2761
8	2	841,7136	83.33	3	91.2281 (52/57)	86.6667 (13/15)	7.1962
9	3	4937,4982,7136	89.0351	4	96.4912 (55/57)	93.3333 (14/15)	7.2058
10	2	7136,4341	92.1053	4	92.9825 (53/57)	86.6667 (13/15)	7.1477
Avg.	3.3		87.9383		95.0877	90.6667	7.2146
Std. Dev.	0.8233		5.9828		3.3898	4.6613	0.0851

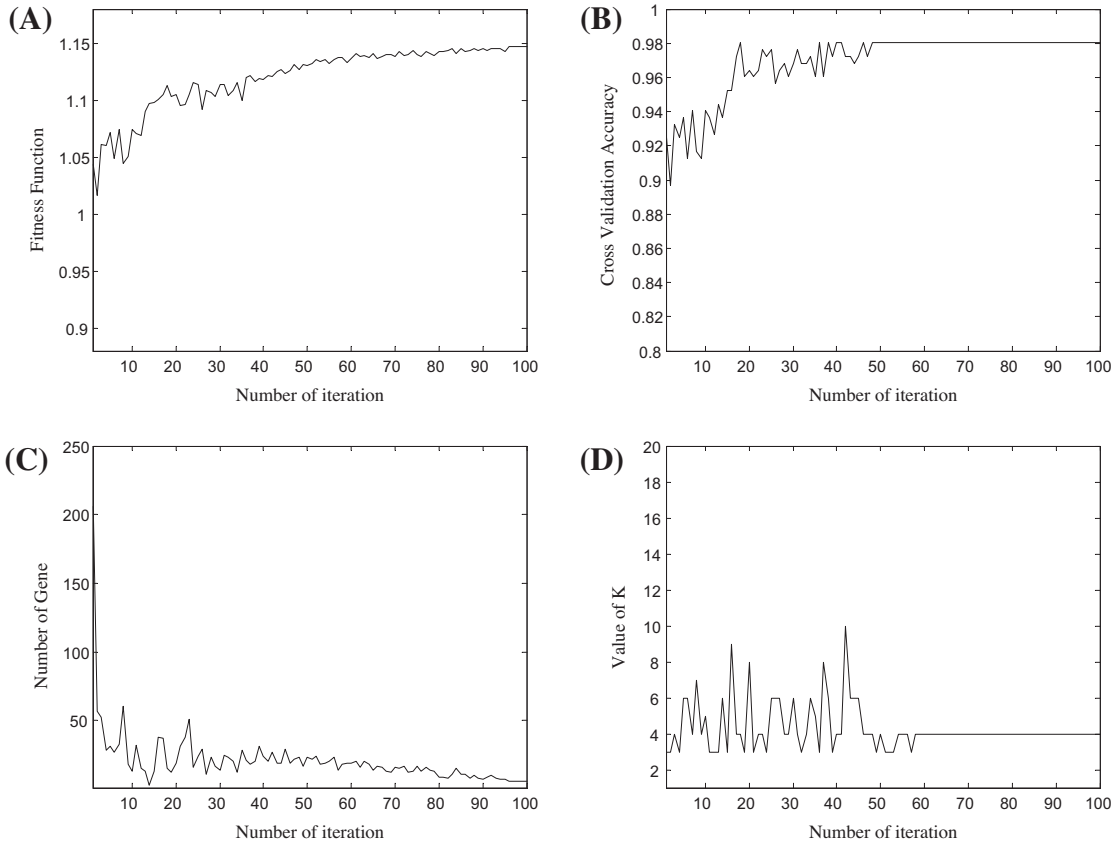

Fig. 1. Result after 100 PSO iterations on SRBCT data (A) fitness function is increasing with number of iterations, (B) number of gene is decreasing with number of iterations, (C) cross validation accuracy initially increases with number of iterations and then gets saturated, and (D) average value of *K* remains almost constant and finally gets saturated at a value of 4.

the classes of different subgroups of cancer. Moreover, these redundant genes increase the dimensionality of the gene subset. In contrast, wrapper methods are usually superior to filter methods since they measure the degree of inter-correlation amongst individual the genes (Chen & et al., 2014).

In recent years, several PSO-based gene selection methods (Li et al., 2008; Shen et al., 2008) and design of classifiers with meta-heuristics solutions (Castillo, Melin, Ramirez, & Soria, 2012; Melin & Castillo, 2013, 2014; Melin et al., 2013) have been explored. The PSO-based gene selection methods have used recently to identify important genes (Li et al., 2008; Shen et al., 2008) because PSO, as an optimizing meta-heuristic tool, embodies a very simple concept and requires only primitive mathematical operators. In addition to that, it is computationally inexpensive in terms of both memory requirements and speed (Eberhart & Kennedy, 1995). None the less, these PSO-based methods usually provide good accuracy while optimizing correlations among the genes (Zheng, Huang, Zhang, & Kong, 2009). Castillo et al. (2012) proposed a hybrid intelligent system for classification of cardiac arrhythmias. They have used three methods of classification, namely, fuzzy *K*-nearest neighbors, multi-layer perceptron with gradient descent and momentum back-propagation algorithm, and multi-layer perceptron with scaled conjugate gradient back-propagation algorithm. Again, Melin & Castillo, 2013 proposed an improvement to the convergence and diversity of the swarm in PSO using fuzzy logic.

A hybrid PSO and tabu search (TS) based approach have been applied by Shen et al., 2008 to select a subset of genes for classification of cancers using gene expression data. At the onset, they have applied a gene ranking method, t-test to rank the genes, and then they have utilized a heuristic method to the microarray datasets. The results obtained in this method have been further improved by Li et al., 2013 with a proposition of GA search space splicing PSO to locate the global optima in the subspaces. Li et al., 2008 have used a Wilcoxon rank sum test to find a crude gene subset on the training samples. Then they have run a hybrid PSO–GA algorithm on the crude gene subset. The accuracy of classification is not appreciable enough and too many genes are selected in the subset for classification of cancers. Chen et al., 2014 have proposed a PSO-based method combined with C4.5 decision tree classifier. They have computed 5-fold cross validation accuracy on the datasets. Mohamad et al., 2011 have proposed an improved binary PSO to select a small subset of informative genes for cancer classification. They have evaluated leave-one-out-cross-validation (LOOCV) accuracy on the datasets though this method is severely constrained by the computation time. The blind test accuracy evaluations on the datasets were not performed in this method.

To summarize and to the best of our knowledge, other researchers (Li et al., 2008; Mohamad et al., 2011; Shen et al., 2008) have attempted to decrease the dimensionality of the dataset by applying a gene ranking method such as t-test (Shen et al., 2008), Wilcoxon rank sum test (Li et al., 2008) and gain ratio (Mohamad et al., 2011) methodologies. Then only they have applied some heuristic methods to this reduced number of genes for exact categorization of a particular type of carcinoma. Therefore, they have considered only a few numbers of genes for the heuristic methods whereas large numbers of genes are completely left out. Moreover, the lacunae that could be identified in their work are enumerated as follows. The blind testing on the datasets was needed to evaluate the classification ability of the selected genes. Further, the usefulness of the selected genes was supposed to be reconfirmed by using a different classifier.

Thus, the basic aim of the proposed work is to reduce the number of genes, particularly suitable for identification of particular type of carcinoma. The number of identifier genes has been attempted to minimize employing the stochastic optimization technique, namely PSO. For that, we have allowed PSO to search the complete gene space defined by the dataset of microarray. Next, we have used several classifier tools including SVM, KNN so that types of carcinoma could be categorized. Here it is mention worthy that compared to the earlier works done by Shen et al. (2008) and others, our PSO-based effective gene detection mechanism is quite simple and it does not use the gene ranking methods to decrease the dimensionality of the datasets. Instead, our proposed method associates a common threshold value to each candidate solution vector and depending upon that threshold value the corresponding gene is retained in the pool; otherwise the gene is discarded. That means, the proposed method pokes all the genes collectively and picks up the effective genes and subsequently KNN and SVM classifiers were run to identify the cancer sub classes based on these selected genes. The present work finally rests on KNN classifier as after performance comparison KNN emerges to be the superior one compared to the SVM. The rationale behind the use of KNN as classifier is that it is non-parametric, lazy (Zhang & Zhou, 2007) and gives better performance in many cases *vis-a-vis* much complicated classifiers proposed in Dudoit, Fridlyand, & Speed, 2002; Vandeginste et al., 1998. However, for the KNN based multi-class classification, the challenge is to choose a proper value of *K* that would guide the search in the appropriate number of neighborhoods and consequently would influence the accuracy of the classified data set. It has been observed that if the value of *K* is less than 2 then the resulting classifier may lack the essential robustness (Tarlow, Swersky, Charlin, Sutskever, & Zemel, 2013). In these cases, the KNN classifier turns out to be a nearest neighbor classifier that suffers from high degree of local sensitivity and incidentally becomes highly susceptible to the noise present in the training data. More robust classifier model is realizable if the value of *K* is chosen to be greater 2, where the majority votes decide the outcome of the class label. A higher value of *K* also results in a smother, less locally sensitive function (Ougiaroglou, Nanopoulos, Papadopoulos, Manolopoulos, & Welzer-Druzovec, 2007). However, if the value of *K* is too high it may yield poor classification results (Ougiaroglou et al., 2007). Generally, KNN classifiers are devised presuming a pre-fixed value of *K*. In contrast, in this paper first we have proposed a heuristic (*K*opt_decision_heu) for judiciously selecting the values of *K* that would result in a classifier with optimum accuracy yet having less computational overhead. Moreover, the PSO-based gene selection heuristic (PSO_gene_select) is interleaved with *K*opt_decision_heu in a manner that the selection process as well as the classification process can adaptively evolve with the values of *K* towards satisfying the goal of attaining better and better gene identification

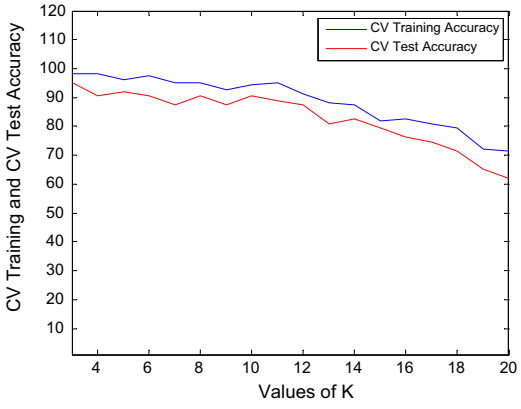

Fig. 2. CV training and CV test accuracy vs. values of *K* for SRBCT dataset.

strategy followed by the goal of obtaining best possible classification accuracy. To the best of our knowledge, till date none of well cited literature has conceived the problem of gene selection and thereafter classification of cancer sub-type from the genes in this orientation. The results also corroborate the fact that our interleaved heuristical approach can reduce the number of essential yet non-redundant genes significantly and the classification based on adaptive KNN can also improve the classification accuracy appreciably. In essence, this work focuses on finding a small subset of informative genes from the training data to achieve high classification accuracy. Next, the subset of informative genes are applied to blind test samples, which previously were not a part of training samples so that the classification ability of the selected subset of genes could be proved. To conclude, we reiterate that the fitness function of PSO has been devised in such a way so that it can improve the classification accuracy even with the minimum number of genes in the subset by selecting the optimized value K (K_{opt}) i.e., the number of neighbors for KNN classifier. To demonstrate the usefulness of the proposed technique of meaningful gene selection and subsequent classification of different cancer sub-groups three standard benchmark microarray gene expression datasets have been utilized (Armstrong et al., 2002; http://research.nhgri.nih.gov/microarray/Supplement/; http://www.broadinstitute.org/cgi-bin/cancer/datasets.cgi). These datasets include both binary-classes and multi-classes data sets used by other researchers too (Bhattacharyya et al., 2003; Chandra & Gupta, 2011; Fu & Fu-Liu, 2005; Ganesh Kumar, Aruldoss Albert Victoire, Renukadevi, & Devaraj, 2012; Ji, Yang, & You, 2011; Khan et al.,

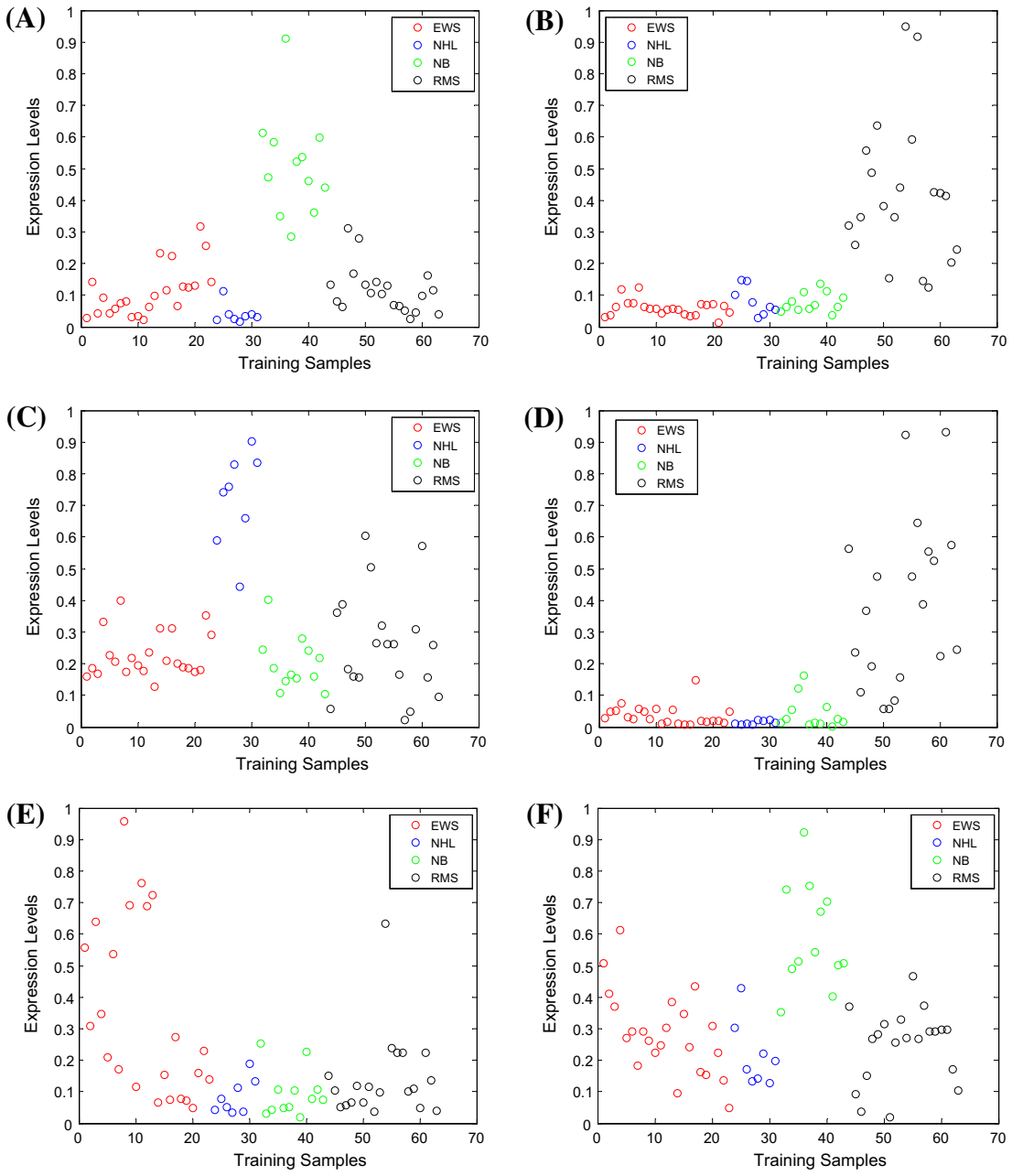

Fig. 3. Scatterplot of six identified genes in the training data of SRBCT dataset. Each panel corresponds to one gene. The red, blue, green and black colors respectively correspond to EWS, NHL, NB and RMS type of SRBCTs. (A) AF1Q is highly expressed for NB, (B) SGCA is highly expressed for RMS, (C) EHD1 is highly expressed for NHL and moderately expressed for a few cases of RMS, (D) EST is highly expressed for RMS, (E) Image ID 234376 is highly expressed for EWS, and (F) Image ID 244637 is highly expressed for NB. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

2001; Lee, Lin, Chen, & Kuo, 2011; Li & Shu, 2009; Mohamad et al., 2011; Pal et al., 2007; Sharma, Imoto, & Miyano, 2012; Shen et al., 2008; Tibshirani, Hastie, Narasimhan, & Chu, 2002; Wong & Liu, 2010; Yang, Cai, Li, & Lin, 2006; Zainuddin & Ong, 2011).

To judge our propositions stated so far, we have relied on the following fact. If the selected subset of genes is essential and sufficient then it should have some universal characteristics. Hence, any standard classifiers should be able to discriminate amongst the genes in the gene subset and consequently should be able to recognize cancer subtypes too. To asses this universal character, we have evaluated the classification accuracies using SVM classifier with different kernels. So to conclude, this paper presents a PSO-based gene selection interleaved with adaptive KNN classification algorithms.

Last but not the least, we have observed the fact that the best combination of gene selection and classification was understood poorly because of the problem of over-fitting i.e. one can obtain good performance using training samples, but when blind test samples is used, a satisfactory result cannot be obtained using the trained model (Horng et al., 2009). However, in this work we have endeavored to obtain the best combination of genes and classification accuracy. Here, we have shown that our propositions and solution methodologies can performance satisfactorily using training samples. And at the same time when blind test samples are used, we have managed to obtain pleasingly agreeable results. This way we have struck the balance between the above mentioned achievable goals.

Finally to conclude this work presents an adaptive meta-heuristic based approach towards minimum number of gene selection for cancer subgroup classification and reconfirmation of the efficacy of the selected genes using rigorous training and testing. Our propositions also ensure high classification accuracy. To the best of our knowledge, such an integrated study in this domain has not been pursued yet.

The paper is organized as follows: Section 2 describes the proposed PSO–adaptive KNN-based gene selection method, Section 3 describes the microarray datasets used for classification purpose and the experimental results obtained by our proposed method, Section 4 describes the discussion on obtained results and finally Section 5 concludes the paper.

2. PSO–adaptive KNN-based gene selection method

2.1. Particle swarm optimization technique

PSO is a stochastic, population-based evolutionary algorithm particularly suitable for solving multi-variable optimization problems. It embeds a kind of swarm intelligence that is based on socio-psychological principles and provides insights into social behavior contributing to engineering applications (Kennedy & Eberhart, 1995). Population dynamics as defined in classical PSO simulates bio-inspired behavior i.e. a bird flock's behavior, which involves sharing of information and allows particles to gain profit from the discoveries and previous experience while in quest of food. In PSO-based applications, each particle represents a candidate solution and flies through the search space. The position of a particle is biased by the best position visited using its own

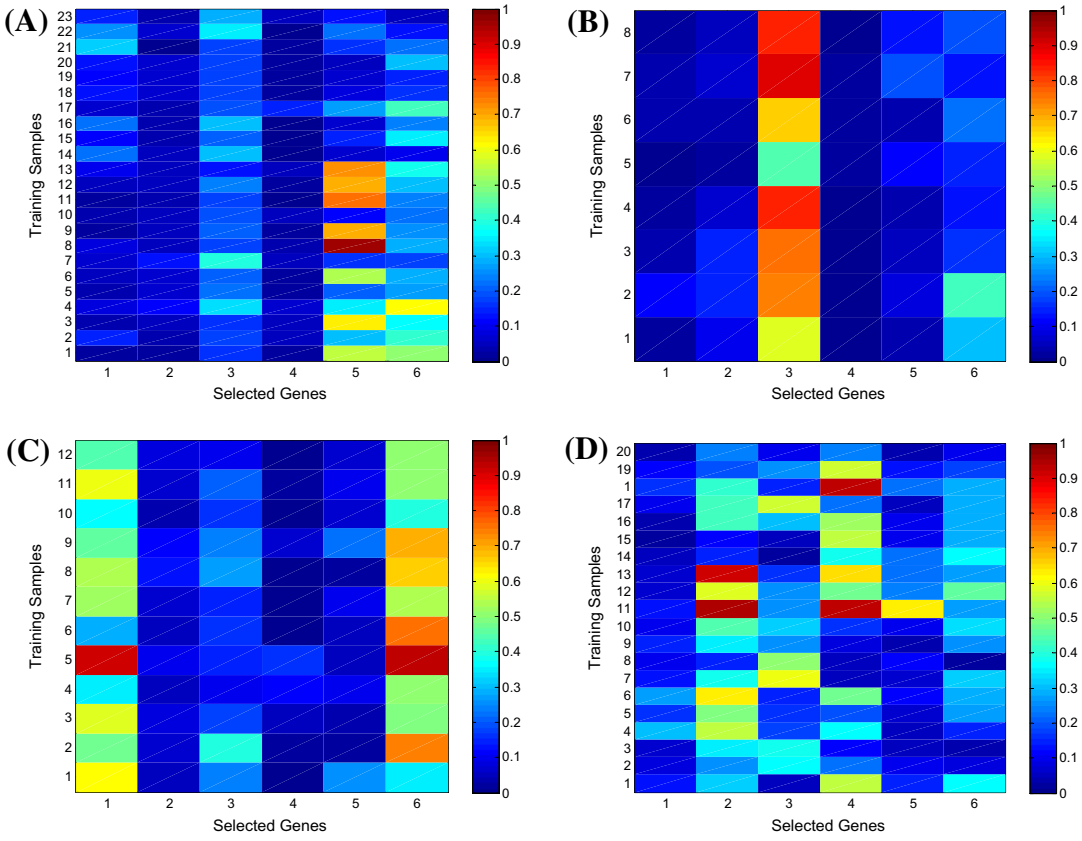

Fig. 4. Pseudo color image of the training data for the 4 SRBCT classes. Each of the four panels represents one class. Transition from blue to red colors represents to a shift from low to high expression levels of the training samples. (A) EWS: The image revels that moderate to high upregulation of Image ID 234376 and moderate upregulation of Image ID 244637 and downregulation for other four genes can signal EWS, (B) NHL: High upregulation of EHD1 and downregulation for other five genes can indicate the presence of NHL, (C) NB: It suggests that high upregulation of AF1Q and Image ID 244637 and downregulation for other four genes are indicator of NB, and (D) RMS: Moderate to high upregulation of SGCA and EST and moderate upregulation for a few cases of EHD1 and Image ID 244637 and downregulation of other two genes can indicate the presence of RMS. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

knowledge and the position of the best particle regarded by the knowledge of neighboring particles. When the neighborhood of a particle is the entire swarm, the particle is said to be the global best particle. How close a particle is to the global optimum is measured by a fitness function, which varies depending upon the optimization problem (Trelea, 2003).

Each particle in the swarm is represented by the following uniqueness:

x_i: current position of the ith particle,
v_i: current velocity of the ith particle,
p_i: best previous position of the ith particle,
gbest: global best particle in its neighborhood.

The personal best position of particle i is the best position experienced by the particle so far. If f is the objective function, the personal best of a particle, at time step Δt is calculated as:

$$p_i(\Delta t+1)=\begin{cases} p_i(\Delta t) & \text{if } f(x_i(\Delta t+1)) \geqslant f(p_i(\Delta t)) \\ x_i(\Delta t+1) & \text{if } f(x_i(\Delta t+1)) < f(p_i(\Delta t)) \end{cases} \quad (1)$$

If *gbest* denotes the global best particle, it is given as:

$$gbest(\Delta t) \in \{p_0, p_1, \ldots, p_s\} = \min\{f(p_0(\Delta t)), f(p_1(\Delta t)), \ldots, f(p_s(\Delta t))\} \quad (2)$$

where s is the size of the entire swarm.

The velocity of particle i is updated by:

$$v_{ij}(\Delta t+1) = w v_{ij}(\Delta t) + c_1 r_1 (p_{ij}(\Delta t) - x_{ij}(\Delta t)) + c_2 r_2 (gbest_j(\Delta t) - x_{ij}(\Delta t)) \quad (3)$$

where v_{ij} represents the jth element in the velocity vector of the ith particle, w is the inertia weight, c_1 and c_2 are the acceleration constants, and r_1, r_2 are random numbers.

The position of particle i, x_i is updated as:

$$x_i(\Delta t+1) = x_i(\Delta t) + v_i(\Delta t+1) \quad (4)$$

The PSO updates instantaneous motion of each particle by using (3) and (4) and this process is continued until a specified number of iterations are exceeded. The excellence of each particle is determined by a fitness function, which reflects the optimality of a particular solution. Algorithm 1 summarized the general PSO technique.

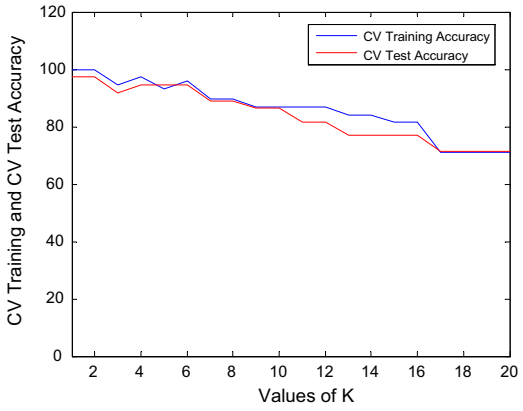

Fig. 6. CV training and CV test accuracy vs. values of K for ALL_AML dataset.

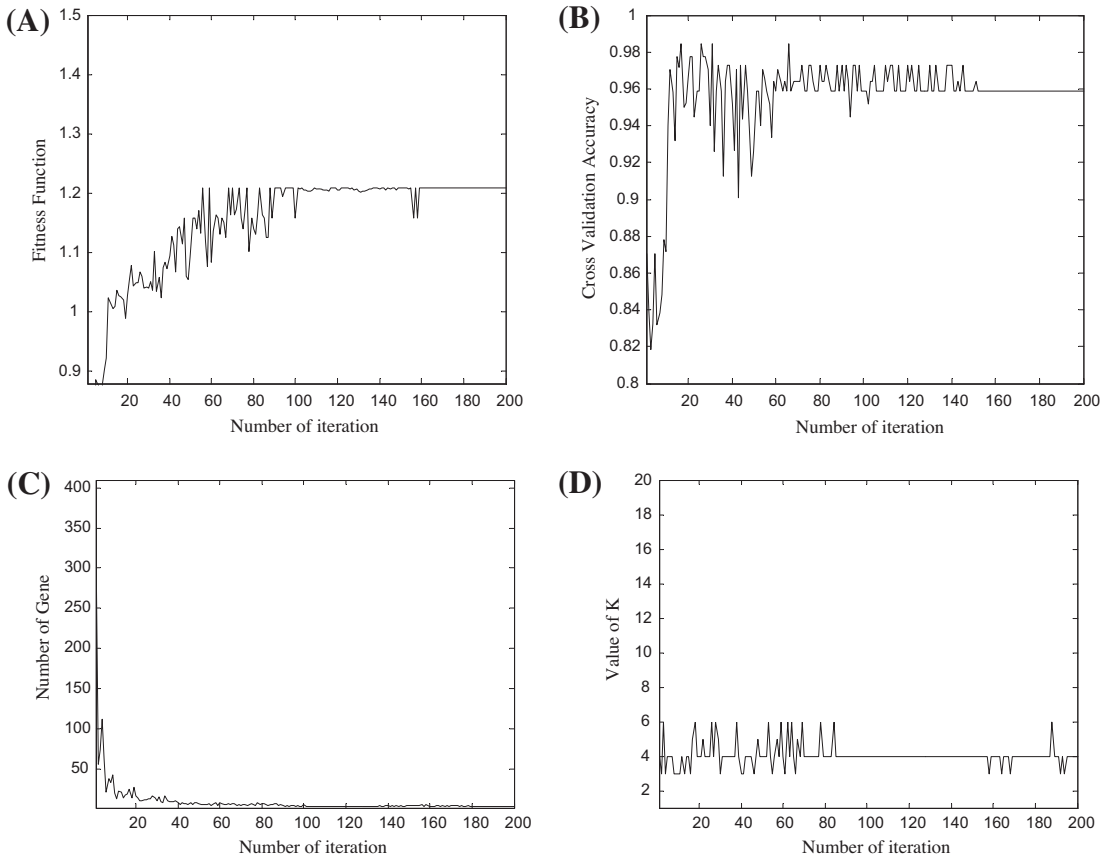

Fig. 5. Result after 200 PSO iterations on ALL_AML data. (A) Fitness function initially increases with number of iterations and then gets saturated (B) Number of gene is decreasing with number of iterations, (C) Cross validation accuracy initially increases with number of iterations and then gets saturated, and (D) average value of K remains almost constant and finally gets saturated at a value of 4.

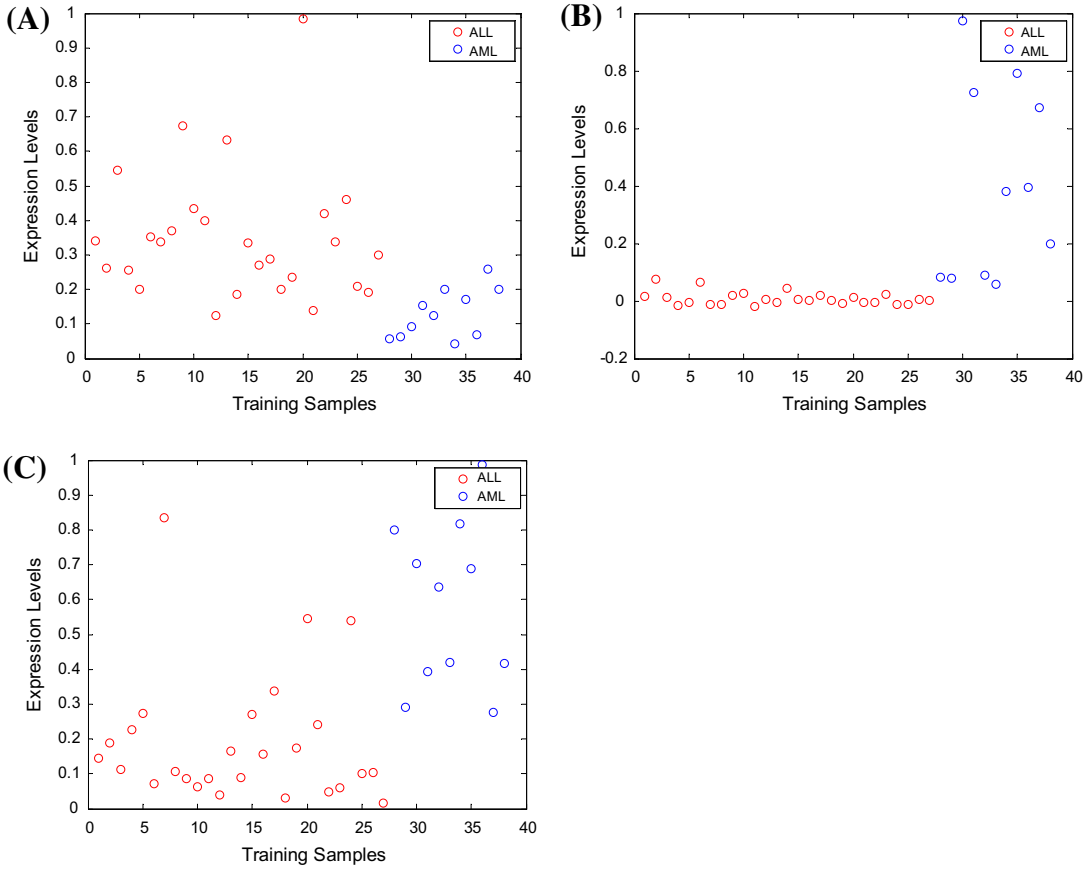

Fig. 7. Scatterplot of three identified genes in the training data of ALL_AML dataset. Each panel corresponds to one gene. The red and blue colors correspond to ALL and AML types respectively. (A) HG1612-HT1612_at is highly expressed for ALL, (B) M27891_at is highly expressed for AML, and (C) X04085_rna1_at is highly expressed for AML and for some cases moderately expressed for ALL. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Algorithm 1: General pseudo code for PSO

for each particle $i \in 1, \ldots, s$ **do**
  initialize position x_i
  initialize velocity v_i
  set $p_i = x_i$
end for
repeat
for each particle $i \in 1, \ldots, s$ **do**
  evaluate the fitness function for each particle i, $f(x_i)$
  evaluate personal best (*pbest*) and global best (*gbest*)
  for each dimension $j \in 1, \ldots, d$ **do**
    calculate new velocity $v_{ij}(\Delta t + 1)$ using Eq. (3)
  end loop
  calculate new position $x_i(\Delta t + 1)$ using Eq. (4)
end loop
until some convergence criteria is satisfied.

2.2. Proposed PSO–adaptive KNN-based gene selection technique

2.2.1. Adaptive K-nearest neighborhood

If the entire data sample G is expressed as $\{G[i, j]\}$, where i is the number of genes and j is the number of samples, a k-fold cross validation technique (Simon, Subramanian, Li, & Menezes, 2011) is employed to divide the entire data sample into k equal subsets of samples. Of the k subsets of samples, a single subset of samples is retained as the validation data for testing and the remaining $(k-1)$ subsets of samples are used as training data. This process then repeated k times, with each of the k subset of samples used exactly once as the validation data. Each time value of K for KNN classifier is varied from 3 to 20.

Cross validation training accuracy (CV training accuracy) and cross validation test accuracy (CV test accuracy) are calculated as:

$$\text{CV training accuracy} = \frac{1}{k}\sum_{m=1}^{k} Tr_m \tag{5}$$

$$\text{CV test accuracy} = \frac{1}{k}\sum_{m=1}^{k} Ts_m \tag{6}$$

where Tr and Ts are the training accuracy and test accuracy for each fold respectively.

For each K, K numbers of single estimations (SE) are calculated as:

$$SE(K) = \frac{1}{2}\sum_{K=3}^{20}\left(\frac{1}{k}\sum_{m=1}^{k} Tr_m + \frac{1}{k}\sum_{m=1}^{k} Ts_m\right) \tag{7}$$

From that K number of single estimations an optimal value is selected. In our case, the optimal value is chosen as: max $\{SE(K)\}$. This optimal value of single estimation is named as cross validation accuracy (CV accuracy) and the corresponding K value for KNN classifier is selected as K_{opt}. Selection of K_{opt} and calculation of CV accuracy are computed for each particle for all the iterations of PSO–adaptive KNN algorithm. We have applied 3-fold cross validation in this present work to reduce the computational time.

2.2.2. Fitness function

In the proposed PSO–adaptive KNN-based gene identification technique, the initial positions of the particles are taken as random numbers between 0 and 1. A threshold ε is utilized to construct the gene identification flags, which will mark the genes i.e. in particular iteration genes are identified from the gene expression dataset, for all training samples, whose corresponding particle position is greater than or equal to the threshold value ε and a candidate gene set is evolved. This candidate gene set is used to calculate the CV accuracy. The fitness function (F) for the PSO–adaptive KNN operation is chosen as:

$$F = A_c + \frac{1}{|N_G|} \tag{8}$$

where A_c = CV accuracy, $|N_G|$ = number of genes in the candidate gene set.

Proposed PSO–adaptive KNN-based gene selection algorithm for classifying the cancer sub classes from microarray gene expression data is described in Algorithm 2.

Algorithm 2: Proposed PSO–adaptive KNN-based gene selection algorithm

for each particle $i \in 1, ..., s$ **do**
  initialize position x_i
  initialize velocity v_i
  set $p_i = x_i$
end for
repeat
  // gene selection flag //
  for each particle $i \in 1, ..., s$ **do**
    for each dimension $j \in 1, ..., d$ **do**
    training samples, $T = \{G(:, j) | x(i, j) \geqslant \varepsilon\}$
    end for
    repeat
    // selection of K_{opt} and calculation of CV accuracy //
      for each $K \in 3, ..., 20$ **do**
        train the classifier using: $\{T(k-1, j)\}$
        test the classifier using: $\{T(k, j)\}$
      end for
      until each of the k fold used once as a test fold.
      for each $K \in 3, ..., 20$ **do**
      evaluate CV training accuracy using Eq. (5)
      evaluate CV test accuracy using Eq. (6)
      evaluate $SE(K)$ using Eq. (7)
       end for
    evaluate CV accuracy as: max $\{SE(K)\}$
    select the value of K_{opt}
    select $|N_G|$ corresponding to the CV accuracy
  // fitness function //
  evaluate the fitness function: $F = A_c + \frac{1}{|N_G|}$
  // personal best & global best particle //
  evaluate personal best ($pbest$)
  end for
  evaluate global best ($gbest$)
  // update position & velocity //
  calculate new velocity and position using Eqs. (3) and (4)
until convergence criteria is satisfied

3. Experimental study

3.1. Data sets

In this present work, three benchmark gene expression datasets (Armstrong et al., 2002; http://research.nhgri.nih.gov/microarray/Supplement/; http://www.broadinstitute.org/cgi-bin/cancer/datasets.cgi) including binary class and multi-class with thousands of genes has been used. The basic information about the datasets are described as follows:

SRBCT Data (http://research.nhgri.nih.gov/microarray/Supplement/): This dataset contains four classes of cancers, namely Ewing sarcoma (EWS), non-Hodgkin lymphoma (NHL), neuroblastoma (NB) and rhabdomyosarcoma (RMS). There are total 2308 number of genes and 88 samples of which 63 samples (EWS:23, NHL:8, NB:12, RMS:20) are used for training. Remaining 25 samples include 5 non SRBCT samples. So for blind testing of the system, 20 samples (EWS:6, NHL:3, NB:6, RMS:5) are used (Pal et al., 2007).

ALL_AML Data (http://www.broadinstitute.org/cgi-bin/cancer/datasets.cgi): This dataset contains two classes of cancers, namely acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML). There are 7129 number of genes with a training dataset containing 38 samples (ALL:27, AML:11) and a test data containing 34 samples (ALL:20, AML:14).

MLL Data (Armstrong et al., 2002): This dataset contains three classes of cancers, namely acute lymphoblastic leukemia (ALL), myeloid lymphoid leukemia (MLL) and acute myeloid leukemia (AML). There are 12,582 numbers of genes. The training set contains 57 samples (ALL:20, MLL:17, AML:20) and test set contains 15 samples (ALL:4, MLL:3, AML:8).

3.2. Experimental results

A number of experiments are separately conducted ten times on each dataset using the proposed method. Next, an average result of these ten independent evaluations is obtained. The average results are important because the proposed method is a stochastic approach. To evaluate the performance of the proposed technique, five important criteria are considered: (a) number of informative genes in the subset, (b) CV accuracy, (c) training accuracy on training samples, (d) test accuracy on blind test samples, and (e) computation time. High accuracy for blind test, small number of selected genes in the subset and lower computation time are essential to achieve excellent performance.

Experimental results for each run on the three datasets are given in the Tables 1–3 respectively. Based on the average results, less than ten numbers of genes are selected to get more than 90% blind test accuracy on the three data sets. However, for each dataset, results of the best subsets shown in shaded cells. This shows that a near optimal gene subset from high dimensional gene expression data has efficiently been selected by the proposed methodology.

Relation of CV accuracy, number of genes and fitness with the number of iterations for the three datasets are shown in Figs. 1, 5 and 9. CV accuracy increases with the number of iterations whereas number of gene decreases with number of iterations. As fitness value is obtained by combining the CV accuracy and number of genes, hence it also increases with the progress of iterations for all the datasets. However, in this work, equal importance is given to the CV accuracy and the number of genes. This trend indicates that the proposed method is appropriate to find a near optimal gene subset from dimensionally large microarray gene expression data with high classification accuracy on blind test samples. However, the numbers of iterations to reach a good solution are problem dependent (Engelbrecht, 2005). Therefore, different numbers of iterations are chosen for three different datasets to obtain near optimal solutions in terms of high blind test accuracy with less number of genes from the gene expression data.

The CV training accuracy and CV test accuracy corresponding to the optimum fitness value for SRBCT, ALL_AML and MLL datasets are shown in Figs. 2, 6 and 10 respectively. We observe that the value of CV accuracy is maximum for $K_{opt} = 4$, where K_{opt} is the

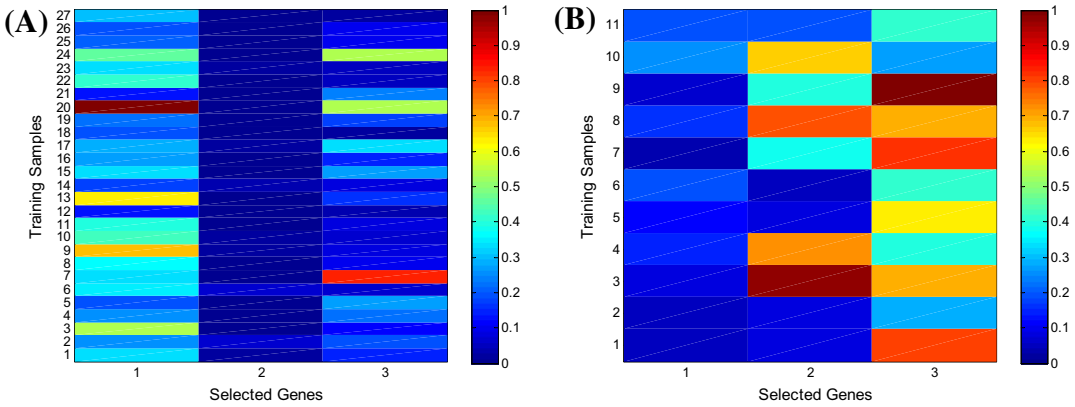

Fig. 8. Pseudo color image of the training data for the 2 ALL_AML classes. Each of the two panels corresponds to one class. Blue to red colors represents low to high expression levels of the training samples. (A) ALL: The image revels that moderate to high upregulation of HG1612-HT1612_at and moderate upregulation of X04085_rna1_at and downregulation for other gene can signal ALL and (B) AML: High upregulation of X04085_rna1_at and moderate to high upregulation for a few cases of M27891_at and downregulation for other gene can indicate the presence of AML. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

number of neighbors for adaptive KNN classifier, for all the three datasets. The scatter plots for all the selected genes for the three datasets SRBCT, ALL_AML and MLL are shown in Figs. 3, 7 and 11 respectively. Each panel corresponds to one gene and the expression levels of the different classes for all the selected genes for all three datasets are shown. The expression values of the training samples of a particular class for the three datasets are shown in Figs. 4, 8 and 12 respectively for each datasets, where transition from blue to red colors corresponds to a shift from low to high expression values of the training samples for each class of all the three datasets.

4. Discussion on obtained results

To select a small and discriminative subset of genes from tens of thousands of genes is extremely hard. Therefore, gene selection

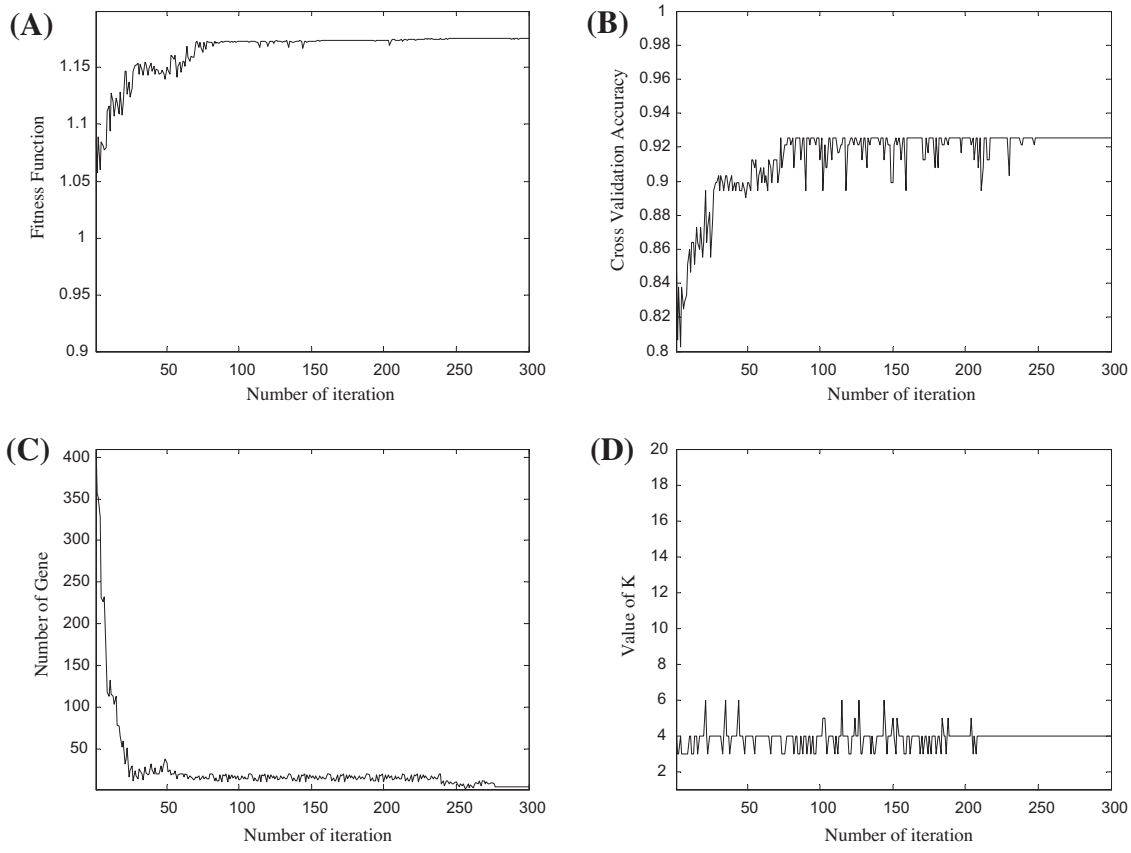

Fig. 9. Result after 300 PSO iterations on MLL data. (A) Fitness function initially increases with number of iterations and then gets saturated, (B) Number of gene is decreasing with number of iterations, (C) Cross validation accuracy initially increases with number of iterations and then gets saturated, and (D) average value of K remains almost constant and finally gets saturated at a value of 4.

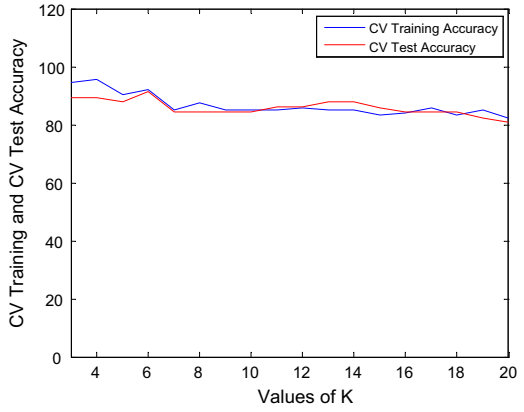

Fig. 10. CV training and CV test accuracy vs. values of *K* for MLL dataset.

becomes the most needed requirement for a diagnostic classifying system. For that reason, many researchers have applied different techniques to select a small subset of informative genes that can classify different subgroups of cancers accurately. A comparison between different methods on the SRBCT, ALL_AML and MLL datasets has shown in Tables 4–6 respectively.

An artificial neural network based method was applied by Khan et al. (2001) and 96 genes were found as important to classify the SRBCT test samples with 100% accuracy. A nearest shrunken centroid based method was applied by Tibshirani et al. (2002) and 43 genes were found as important with 100% accuracy on training and test samples of SRBCT. A feature selection multi layered perceptron and non-Euclidean relational fuzzy c-means clustering based method was applied by Pal et al. (2007) and only 7 genes were identified as important for 100% classification accuracy on training and test data of SRBCT. Lee et al. (2011) have used an adaptive genetic algorithm/KNN based method to evolve gene subsets. They have found only 14 genes to classify test samples of SRBCT dataset with 100% accuracy.

An algorithm based on support vector machine (SVM) and recursive feature elimination was adopted by Fu and Fu-Liu (2005) and applied on both SRBCT and ALL_AML dataset. 19 genes and 4 genes were identified as important for SRBCT and ALL_AML dataset respectively. 100% classification accuracy on the test samples of SRBCT dataset and 97.06% classification accuracy on the test samples of ALL_AML dataset were achieved by their proposed method. Shen et al. (2008) have applied four techniques such as stepwise, pure tabu search, pure PSO and hybrid PSO–TS and identified 3, 5, 7 and 7 genes respectively for ALL_AML data with 88.14%, 94.24%, 94.19% and 95.81% accuracy on test samples respectively. Ji et al. (2011) proposed two methods namely partial least square variant importance in projection (PLSVIP) and partial least square independent-variable explanation gain (PLSIEG) for classification of SRBCT and ALL_AML datasets. For SRBCT dataset, 24 genes were identified for 100% classification accuracy on test samples, and for ALL_AML dataset, 9 genes were identified for 100% classification accuracy on test samples, by applying PLSVIP technique. Whereas for SRBCT data 15 genes and for ALL_AML data 8 genes were identified to achieve 100% classification accuracy on test samples by applying PLSIEG technique. Ganesh Kumar et al.

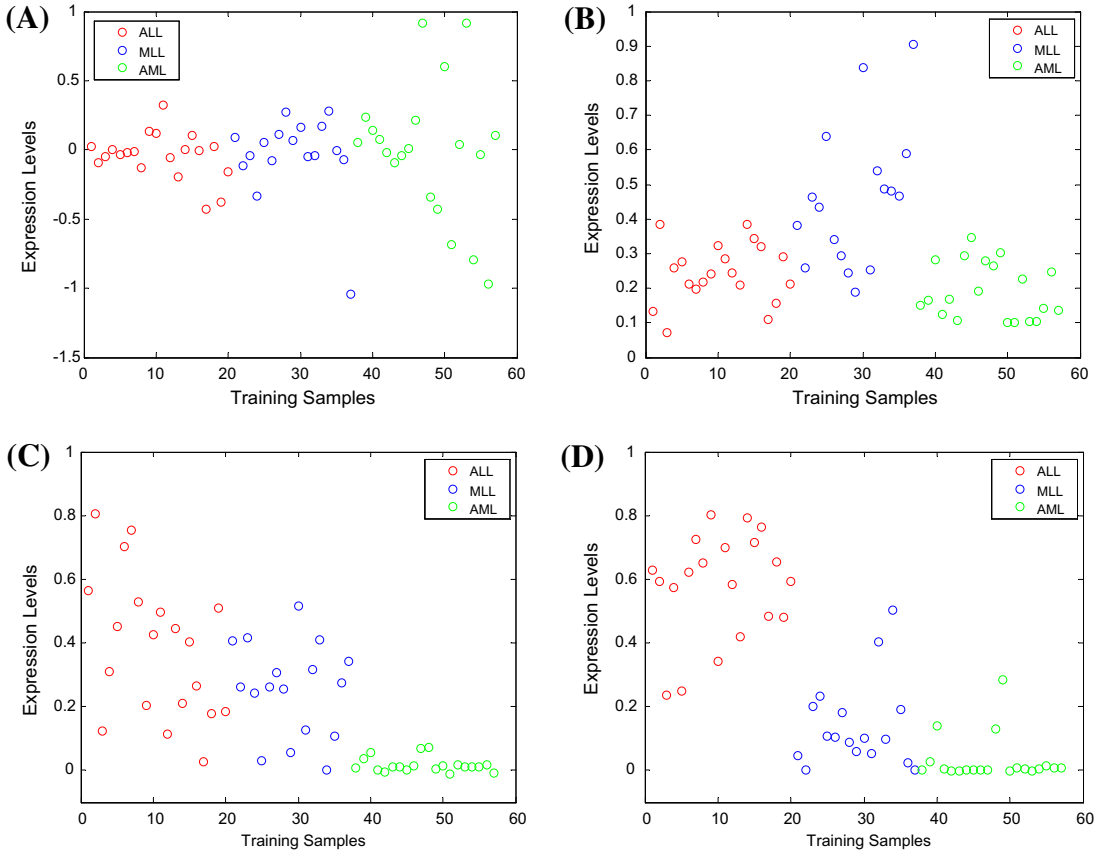

Fig. 11. Scatterplot of four identified genes in the training data of MLL dataset. Each panel corresponds to one gene. The red, blue and green colors correspond to ALL, MLL and AML types respectively. (A) 33054_at is moderate to highly expressed for AML, (B) 34306_at is highly expressed for MLL, (C) 37710_at is moderate to highly expressed for ALL and for a few cases of MLL, and (D) 266_s_at is highly expressed for ALL and moderately expressed for a few cases of MLL. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

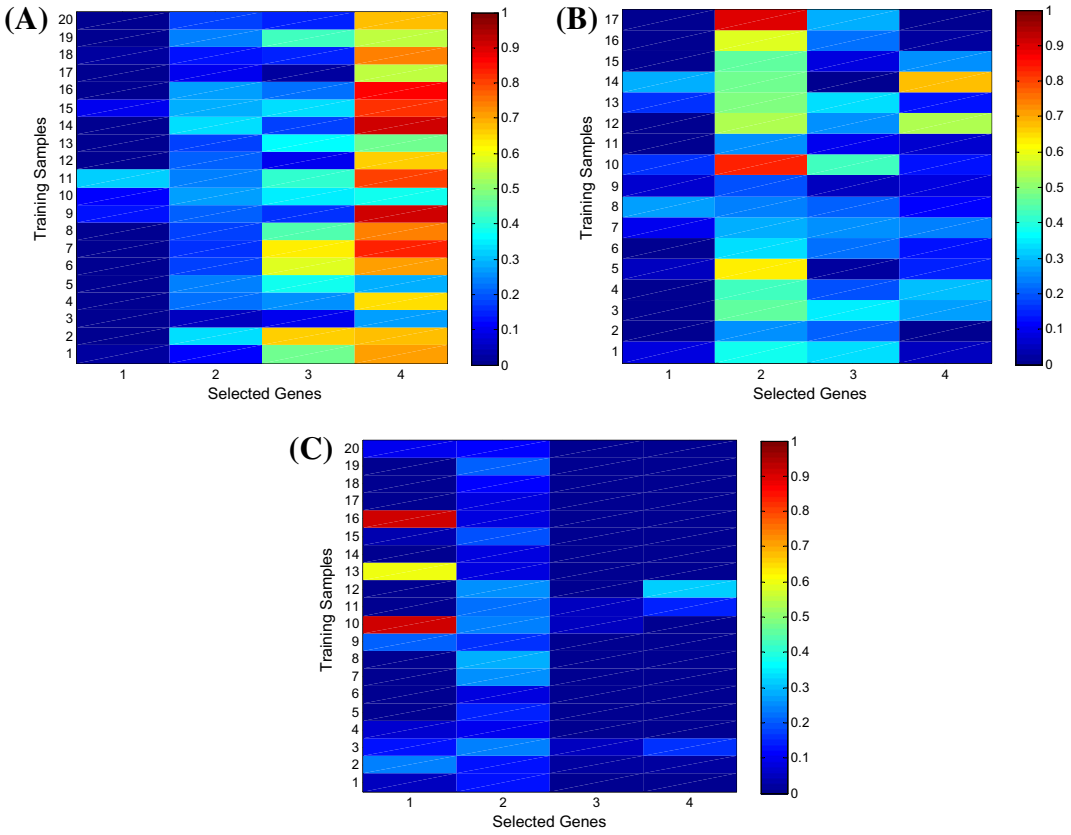

Fig. 12. Pseudo color image of the training data for the 3 MLL classes. Each of the three panels corresponds to one class. Blue to red colors represents low to high expression levels of the training samples. (A) ALL: The image revels that high upregulation of 266_s_at and moderate to high upregulation of 37710_at and downregulation for other two genes can signal ALL, (B) MLL: High upregulation of 34306_at and moderate to high upregulation for a few cases of 37710_at and downregulation for other two genes can indicate the presence of MLL, and (C) AML: It suggests that moderate to high upregulation for a few cases of 33054_at and downregulation for other three genes are indicator of AML. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 4
Comparison of the methods on the SRBCT data set.

Experiments	Methods	Genes				Cross validation accuracy (%)				Training accuracy (%)	Test accuracy (%)
Khan et al. (2001)	ANN	96				–				100	100
Tibshirani et al. (2002)	NSC	43				–				100	100
Fu and Fu-Liu, (2005)	SVM-RFE	19				–				100	100
		5CV		LOOCV		5CV		LOOCV		–	–
		KNN	SVM	KNN	SVM	KNN	SVM	KNN	SVM	–	–
Yang et al. (2006)	GS1	88	93	57	34	98	97.9	98.8	98.8		
	GS2	90	99	77	96	98.1	99	98.8	100	–	–
	Cho's	98	98	82	80	90.2	94.3	92.8	98.8	–	–
	F-test	90	95	89	78	98	99.2	98.8	100	–	–
Pal et al. (2007)	FSMLP + NERFCM	7				–				100	100
Ji et al. (2011)	PLSVIP	24				–				100	100
	PLSVEG	15				–				100	100
Mohamad et al. (2011)	IBPSO	6				100				–	–
Sharma et al. (2012)	SFS + LDA with NCC	4				–				100	100
	SFS + Bayes classifier	4				–				100	90
	SFS + NNC	4				–				100	95
Zainuddin and Ong (2011)	MSFCM + WNN	10				10CV 100				–	–
Li and Shu (2009)	KLLE + LLE + PCA	20				–				100	100
Lee et al. (2001)	AGA + KNN	14				–				100	100
Chen et al. (2014)	PSODT	–				5CV 92.94				–	–
This paper	**Proposed technique**	**6**				**98.0159**				**100**	**100**

Table 5
Comparison of the methods on the ALL_AML data set.

Experiments	Methods	Genes				Cross validation accuracy (%)				Training accuracy (%)	Test accuracy (%)
Fu and Fu-Liu (2005)	SVM-RFE	4				–				100	97.06
		5CV		LOOCV		5CV		LOOCV		–	–
		KNN	SVM	KNN	SVM	KNN	SVM	KNN	SVM		
Yang et al. (2006)	GS1	100	93	60	4	97.9	97.9	98.6	98.6	–	–
	GS2	85	98	10	25	97.1	97.4	98.6	98.6	–	–
	Cho's	100	98	9	80	96.8	97	97.2	98.6	–	–
	F-test	96	99	25	33	97.4	97.5	98.6	98.6	–	–
Shen et al. (2008)	Stepwise	3				–				90.83	88.14
	Pure TS	5				–				95.83	94.24
	Pure PSO	7				–				94.75	94.19
	HPSOTS	7				–				98.08	95.81
Ji et al. (2011)	PLSVIP	9				–				100	100
	PLSIEG	8				–				100	100
Mohamad et al. (2011)	IBPSO	2				100				–	–
Zainuddin and Ong (2011)	MSFCM + WNN	10				10 CV 98.61				–	–
Wong and Liu (2010)	Probabilistic mechanism	–				SVM		KNN		–	–
						97.38		98.21			
Chandra and Gupta (2011)	RNBC	–				10CV				–	–
						RNBC	NBC	KNN			
						94.29	84.29	85.71			
Ganesh Kumar et al. (2012)	GSA	10				100				–	–
This paper	**Proposed method**	**3**				**95.8868**				**100**	**97.0588**

Table 6
Comparison of the Methods on the MLL Data Set.

Experiments	Methods	Genes				Cross validation accuracy (%)				Training accuracy (%)	Test accuracy (%)
Yang et al. (2006)		5CV		LOOCV		5CV		LOOCV		–	–
		KNN	SVM	KNN	SVM	KNN	SVM	KNN	SVM		
	GS1	29	99	97	56	94.8	95.2	97.2	97.2	–	–
	GS2	91	87	90	91	94.9	94.7	97.2	97.2	–	–
	Cho's	93	89	23	44	96	95.5	97.2	95.8	–	–
	F-test	99	100	65	31	95.4	94.8	95.8	95.8	–	–
Sharma et al. (2012)	SFS + LDA with NCC	4				–				100	100
	SFS + Bayes classifier	4				–				100	100
	SFS + NNC	4				–				100	93
Mohamad et al. (2011)	IBPSO	4				100				–	–
Chandra and Gupta (2011)	RNBC	–				10CV				–	–
						RNBC	NBC	KNN			
						87.14	80	68.57			
Chen et al.(2014)	PSODT	–				5CV 100				–	–
This paper	**Proposed method**	**4**				**92.5439**				**100**	**100**

Table 7
List of best subset of genes for SRBCT data.

Index no. of selected genes	Image ID	Gene ID	Gene name
742	812105	AF1Q	Transmembrane Protein
1003	796258	SGCA	Sarcoglycan, alpha(50KD dystrophin-associated glycoprotein)
1386	745019	EHD1	EH domain containing 1
2046	244618	EST	ESTs
2099	234376	–	Homo Sapiens mRNA: cDNA DKFZp564F112(from clone DKFZp564F112)
2157	244637	–	Homo Sapience mRNA full length insert cDNA clone EUROIMAGE

(2012) have proposed a novel Genetic Swarm Algorithm for obtaining near optimal rule set. They have proposed this method for acquisition of knowledge in the form of if–then rules and membership functions. They have found only 10 genes to achieve 100% leave-one-out-cross-validation-accuracy (LOOCV) on ALL_AML data. A modified wavelet neural network based method was applied by Zainuddin and Ong (2011) on both SRBCT and ALL_AML datasets. Only 10 genes were found important to achieve 100% 10-fold cross validation accuracy on SRBCT data and 98.61% 10-fold cross validation accuracy on ALL_AML dataset.

Successive feature selection (SFS) with linear discriminant analysis (LDA) and nearest centroid classifier (NCC), SFS with Bayes classifier and SFS with nearest neighbor classifier (NNC) techniques

Table 8
List of best subset of genes for ALL_AML data.

Index no. of selected genes	Gene accession number	Gene description
804	HG1612-HT1612_at	Macmarcks
1882	M27891_at	CST3 cystatin C (amyloid angiopathy and cerebral hemorrhage)
4052	X04085_rna1_at	Catalase (EC 1.11.1.6) 5'flank and exon 1 mapping to chromosome11, band p13 (and joined cds)

Table 9
List of best subset of genes for MLL data.

Index no. of selected genes	Gene accession number	Gene description
841	33054_at	Hs.144852 gnl|UG|Hs#S1570396 Homo sapiens mRNA; cDNA DKFZp434N074 (from clone DKFZp434N074)
7930	34306_at	Hs.28578 gnl|UG|Hs#S892244 Homo sapiens KIAA0428 mRNA, complete cds
8937	37710_at	Hs.78995 gnl|UG|Hs#S2842 Homo sapiens MADS
12418	266_s_at	Hs.278667 gnl|UG|Hs#S885 Homo sapiens CD24 signal transducer mRNA, complete cds and 3' region

Table 10
Essential and sufficient conditions.

Datasets	SVM kernels	Training accuracy (%)	Test accuracy (%)
SRBCT	Linear	98.4127(62/63)	95(19/20)
	RBF with sigma = 1	98.4127(62/63)	65(13/20)
	Polynomial with order 3	100(63/63)	90(18/20)
	Quadratic	100(63/63)	85(17/20)
ALL_AML	Linear	97.3684(37/38)	97.0588(33/34)
	RBF with sigma = 1	100(38/38)	97.0588(33/34)
	Polynomial with order 3	100(38/38)	94.1176(32/34)
	Quadratic	100(38/38)	97.0588(33/34)
MLL	Linear	100(57/57)	100(15/15)
	RBF with sigma = 1	100(57/57)	100(15/15)
	Polynomial with order 3	100(57/57)	100(15/15)
	Quadratic	98.2456(56/57)	93.3333(14/15)

were proposed by Sharma et al. (2012) and applied on both SRBCT and MLL datasets. For SRBCT dataset 4 genes were identified by all these three methods, but on test samples, 100% classification accuracy achieved by combining SFS and LDA with NCC technique, 90% classification accuracy achieved by combining SFS and Bayes classifier technique and 95% classification accuracy achieved by combining SFS and NNC technique. For MLL dataset 4 genes were also identified by all the three methods proposed by Sharma et al. (2012) and achieved 93% classification accuracy on test samples by combining SFS and NNC techniques whereas achieved 100% classification accuracy by the other two methods. Chandra and Gupta (2011) have proposed a robust Naïve-Bayes Classifier (NBC) based method because NBC is widely used for classification in machine learning. They have found 94.29% 10-fold cross validation accuracy on ALL_AML dataset and 87.14% 10-fold cross validation accuracy on MLL dataset. Chen et al. (2014) have used a PSODT technique and achieved 92.94% and 100% 5-fold cross validation accuracy on SRBCT and Leukemia2 dataset respectively.

Yang et al. (2006) applied Gene scoring technique and F-test methods for the gene selection of SRBCT, ALL_AML and MLL datasets. Two different classifier namely KNN and SVM were used and 5-fold cross validation accuracy and leave one out cross validation accuracy (LOOCV) were measured. The numbers of identified genes by their method are too large. An improved binary PSO (IBPSO) based method were proposed by Mohamad et al. (2011) for the selection of gene subsets for 10 different datasets including SRBCT, ALL_AML and MLL datasets. SVM and LOOCV were used to measure the classification accuracy. Classification accuracy on blind test samples has not been tested. For SRBCT, ALL_AML and MLL datasets they identified 6, 2 and 4 genes respectively to achieve 100% LOOCV accuracy.

In contrast, in this present work, only 6 genes for SRBCT data, only 4 genes for MLL data and only 3 genes for ALL_AML data are identified as important for classification of blind test samples. 100% classification accuracy on the blind test samples is achieved for SRBCT and MLL datasets whereas 97.0588% classification accuracy is achieved on ALL_AML dataset, i.e. only one misclassification occurred out of 34 blind test samples.

Table 7 shows the best subset of selected genes for the SRBCT data by the proposed method. It has been observed that AF1Q is highly expressed for NB. AF1Q was also found by Khan et al. (2001), Pal et al. (2007), Fu and Fu-Liu (2005) and Tibshirani et al. (2002). A search in GEO (Gene Expression Omnibus, NCBI) profiles showed that AF1Q discriminates between NB and Ewing family tumor (GPL96, 211071_s_at (ID_REF), GDS1713). Further, it has been observed that SGCA is highly expressed for RMS. SGCA was also found by Khan et al. (2001) and Pal et al. (2007). A search in GEO profiles showed that SGCA can discriminate RMS from Ewing's sarcoma (GPL91, 38609_at (ID_REF),GDS971). EHD1 is highly expressed for NHL and moderately expressed for few cases of RMS in our case. EHD1 was also found by Pal et al. (2007). They also observed that it is up-regulated in NHL and in a few cases of RMS. GEO profiles showed that EHD1 is moderately expressed for RMS (GPL91,40098_at (ID_REF),GDS971). Again, it has been observed that EST is highly expressed for RMS. EST was also found by Khan et al. (2001). GEO profiles (Gene Expression Omnibus, NCBI) showed that EST is highly expressed for RMS (GPL91,33961_at (ID_REF),GDS971). Furthermore, it has been observed that Image ID 234376 is highly expressed for EWS. Image ID 234376 was also found by Bhattacharyya et al. (2003). They also showed that this gene can discriminate between EWS and other SRBCTs. In this case, it has been observed that Image ID 244637 is highly expressed for NB. Yeo & Poggio, 2001 also found image ID 244637.

Table 8 shows the best subset of selected genes selected genes for the ALL-AML data by the proposed method. It has been observed that HG1612-HT1612_at is highly expressed for ALL. Shen et al. (2008) also found HG1612-HT1612_at. GEO profiles showed that HG1612-HT1612_at is highly expressed for ALL (GPL1261,1435627_x_at (ID_REF),GDS4303). Further, it has been observed that M27891_at is highly expressed for AML. M27891_at was also found by Fu and Fu-Liu (2005) and Ji et al. (2011). GEO profiles showed that M27891_at is highly expressed for AML (GPL8300,39689_at (ID_REF),GDS1059). X04085_rna1_at is found by the proposed method. It has been again observed that X04085_rna1_at is highly expressed for AML and for some cases moderately expressed for ALL. Interestingly, a search in GEO profiles also showed a high expression of X04085_rna1_at for AML (GPL8300,37009_at (ID_REF),GDS1059).

Table 9 shows the best subset of selected genes selected genes for the MLL data by the proposed method. It has been observed that 34306_at is highly expressed for MLL. 34306_at was also found by Sharma et al. (2012). GEO profiles showed that 34306_at is moderate to highly expressed for MLL (GPL8300,34306_at (ID_REF),GDS2729). Again, it has been observed that 37710_at is highly expressed for ALL and for a few cases of MLL. 37710_at

was also found by Mohamad et al. (2011). GEO profiles showed that 37710_at is moderate to highly expressed for ALL and for a few cases of AML (GPL8300,37710_at (ID_REF),GDS1059). It has been also observed that 266_s_at is highly expressed for ALL and moderately expressed for a few cases of MLL. 266_s_at was also found by Mohamad et al., 2011. GEO profiles showed that 266_s_at is highly expressed for ALL (GPL570,266_s_at (ID_REF), GDS2970). 33054_at is found by the proposed method. Furthermore, it has been observed that 33054_at is moderately expressed for AML. A search in GEO profiles also showed a moderate expression for AML (GPL8300,33054_at (ID_REF),GDS1059).

Finally, a subset of selected genes should be able to produce good classification accuracy by using a different classifier. To assess this universal character, the classification accuracy, using the selected subset of genes for all the three datasets are evaluated using SVM classifier with different kernels. The results for the three datasets are shown in Table 10 where it is shown that, the classification accuracy by the selected genes, using different kernels of SVM classifier are promising. This reconfirms the usefulness and universal characteristics of the identified genes. The essential and sufficient conditions are used for a gene subset as a whole and hence there could be other such subsets. This shows that the proposed method is sufficiently able to identify the cancer subgroups in terms of classification accuracy on blind test samples, number of informative genes and computing time. Therefore, biologists can rely on the computational methodology proposed in this work and can save more time since they can directly refer to the subset of selected genes for early detection of cancers because the subset of selected genes has high possibility to classify cancer subgroups accurately.

5. Conclusion

In this paper, a PSO–adaptive KNN-based gene selection method for proper classification of microarray data is proposed. A heuristic (Kopt_decision_heu) for selecting the optimal values of K efficiently, guided by the classification accuracy, is also proposed. In the earlier period, researchers used statistical methods to reduce the dimension of the dataset. After that they have applied some heuristic methods to select subset of informative genes. However, due to the reduced dimension, some informative genes would not take part in the heuristic method. Therefore, the contribution of this research may be interpreted as the proposed gene selection methodology is capable of looking at all genes together and picking up whatever is needed. Moreover, the usefulness and universal character of the selected subsets are reconfirmed by using different kernels of SVM classifier. The results of this study indicate that the proposed PSO–adaptive KNN-based method can be provided as a useful tool for gene selection from gene expression data. The proposed method is also compared to other gene selection methods applied on the same datasets and promising results are obtained. It is observed that this method finds 6, 3 and 4 genes to achieve 100% (20/20), 97.0588% (33/34) and 100% (15/15) blind test accuracies for SRBCT, ALL_AML, and MLL datasets respectively.

One of the most potential areas in applying microarray technology is clinical microbiology. It uses the low or middle density microarrays for the simultaneous assessment of large numbers of microbial genetic objects. The proposed method bespeaks the possibility of developing simplified methods for an easy diagnosis of the cancer subgroups. The possibility would be to design specialized microarray chips for these very purposes. Thereafter, histopathologists can identify the relevant genes efficiently and classify the blind test samples correctly.

Furthermore, to realize the advantages of microarray technology based treatment in cancer diagnosis and classification, the improvement of the ancillary technologies of microarray should be developed. New microarray platforms should come up in tandem with the statistics and software for analysis and data mining technologies. Then only more precise and cheaper simplified technical and analytical procedures will benefit the patient in a large context. In addition to that standardized diagnostic techniques employing microarray data has to be formulated in cooperation with different laboratories working for generating meta-profiles in this direction of cancer research.

References

Armstrong, S. A., Staunton, J. E., Silverman, L. B., Pieters, R., den Boer, M. L., Minden, M. D., et al. (2002). MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nature Genetics, 30*, 41–47.

Bhattacharyya, C., Grate, L. R., Rizki, A., Radisky, D., Molina, F. J., Jordan, M. I., et al. (2003). Simultaneous classification and relevant feature identification in high-dimensional spaces: Application to molecular profiling data. *Signal Processing, 83*, 729–743.

Castillo, O., Melin, P., Ramirez, E., & Soria, J. (2012). Hybrid intelligent system for cardiac arrhythmia classification with fuzzy K-nearest neighbors and neural networks combined with a fuzzy system. *Expert Systems with Applications, 39*(3), 2947–2955.

Chandra, B., & Gupta, M. (2011). Robust approach for estimating probabilistics in Naïve-Bayes classifier for gene expression data. *Expert Systems with Applications, 38*, 1293–1298.

Chen, K.-H. et al. (2014). Gene selection for cancer identification: A decision tree model empowered by particle swarm optimization algorithm. *BMC Bioinformatics, 15*, 49.

Chen, D. C., Liu, Z. Q., Ma, X. B., & Hua, D. (2005). Selecting genes by test statistics. *Journal of Biomedicine and Biotechnology, 2*, 132–138.

Dudoit, S., Fridlyand, J., & Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association, 97*, 77–87.

Eberhart, R., & Kennedy, J. (1995). A new optimizer using particle swarm theory. In *Sixth IEEE International Symposium on micro machine and human science* (pp. 39–46).

Engelbrecht, A. P. (2005). *Fundamentals of computational swarm intelligence*. West Sussex, UK: Wiley.

Fu, L. M., & Fu-Liu, C. S. (2005). Evaluation of gene importance in microarray data based upon probability of selection. *BMC Bioinformatics, 6*, 67.

Furey, T. S., Christianini, N., Duffy, N., Bednarski, D. W., Schummer, M., & Haussler, D. (2000). Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics, 16*(10), 906–914.

Ganesh Kumar, P., Aruldoss Albert Victoire, T., Renukadevi, P., & Devaraj, D. (2012). Design of fuzzy expert system for microarray data classification using a novel genetic swarm algorithm. *Expert Systems with Applications, 39*, 1811–1821.

Horng, J. T., Wu, L. C., Liu, B. J., Kuo, J. L., Kuo, W. H., & Zhang, J. J. (2009). An expert system to classify microarray gene expression data using gene selection by decision tree. *Expert Systems with Applications, 36*, 9072–9081.

http://research.nhgri.nih.gov/microarray/Supplement/.

http://ww.wbroadinstitute.org/cgi-bin/cancer/datasets.cgi.

Ji, G., Yang, Z., & You, W. (2011). PLS-based gene selection and identification of tumor-specific genes. *IEEE Transactions on Systems, Man and Cybernetics – Part C: Applications and Reviews, 41*(6), 830–841.

Kennedy, J., & Eberhart, R. C. (1995). Particle swarm optimization. In *Proceedings of the IEEE international conference on neural networks* (Vol. 4, pp. 1942–1948).

Khan, J., Wei, J. S., Ringner, M., Saal, L. H., Ladanyi, M., Westermann, F., et al. (2001). Classification and diagnostic prediction of cancers using gene expressing profiling and artificial neural network. *Nature Medicine, 7*, 673–679.

Lee, C. P., Lin, W. S., Chen, Y. M., & Kuo, B. J. (2011). Gene selection and sample classification on microarray data based on adaptive genetic algorithm/k-nearest neighbor method. *Expert Systems with Applications, 38*, 4661–4667.

Li, H. et al. (2013). Genetic algorithm search space splicing particle swarm optimization as general-purpose optimizer. *Chemometrics and Intelligent Laboratory Systems, 128*, 153–159.

Li, X., & Shu, L. (2009). Kernel based nonlinear dimensionality reduction for microarray gene expression data analysis. *Expert Systems with Applications, 36*, 7644–7650.

Li, S., Wu, X., & Tan, M. (2008). Gene selection using hybrid particle swarm optimization and genetic algorithm. *Soft Computing, 12*, 1039–1048.

Melin, P., & Castillo, O. (2013). A review on the applications of type-2 fuzzy logic in classification and pattern recognition. *Expert Systems with Applications, 40*(13), 5413–5423.

Melin, P., & Castillo, O. (2014). A review on type-2 fuzzy logic applications in clustering, classification and pattern recognition. *Applied Soft Computing, 21*, 568–577.

Melin, P., Olivas, F., Castillo, O., Valdez, F., Soria, J., Mario, J., et al. (2013). Optimal design of fuzzy classification systems using PSO with dynamic parameter adaptation through fuzzy logic. *Expert Systems with Applications, 40*(8), 3196–3206.

Mohamad, M. S., Omatu, S., Deris, S., & Yoshioka, M. (2011). A modified binary particle swarm optimization for selecting the small subset of informative genes from gene expression data. *IEEE Transactions on Information Technology in Biomedicine, 15*(6), 813–822.

Ougiaroglou, S., Nanopoulos, A., Papadopoulos, A. N., Manolopoulos, Y., & Welzer-Druzovec, T. (2007). Adaptive k-nearest neighbor classification using a dynamic number of nearest neighbors. In Advances in databases and information systems: Vol. 4690 Lecture notes in computer science (pp. 66–82).

Pal, N. R., Aguan, K., Sharma, A., & Amari, S. (2007). Discovering biomarkers from gene expression data for predicting cancer subgroups using neural networks and relational fuzzy clustering. *BMC Bioinformatics, 8*, 5.

Sharma, A., Imoto, S., & Miyano, S. (2012). A top-r feature selection algorithm for microarray gene expression data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics, 9*(3), 754–764.

Shen, Q., Shi, W. M., & Kong, W. (2008). Hybrid particle swarm optimization and tabu search approach for selecting genes for tumor classification using gene expression data. *Computational Biology and Chemistry, 32*, 53–60.

Simon, R. M., Subramanian, J., Li, M. C., & Menezes, S. (2011). Using cross-validation to evaluate predictive accuracy of survival risk classifiers based on high-dimensional data. *Briefings in Bioinformatics, 12*(3), 203–214.

Tarlow, D., Swersky, K., Charlin, L., Sutskever, I., & Zemel, R. S. (2013). Stochastic k-neighborhood selection for supervised and unsupervised learning. In *30th International conference on machine learning, Atlanta, GA, USA* (p. 28).

Tibshirani, R., Hastie, T., Narasimhan, B., & Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *PNAS, 99*(10), 6567–6572.

Trelea, I. C. (2003). The particle swarm optimization algorithm: Convergence analysis and parameter selection. *Information Processing Letters, 85*, 317–325.

Vandeginste, B. G. M., Massart, D. L., Buydens, L. M. C., De Jong, S., Lewi, P. J., & Smeyers-Verbeke, J. (1998). *Handbook of chemometrics and qualimetrics* (Vol. 20B). Berlin: Springer.

Wong, T. T., & Liu, K. L. (2010). A probabilistic mechanism based on clustering analysis and distance measure for subset gene selection. *Expert Systems with Applications, 37*, 2144–2149.

Xiong, M. M., Fang, X. Z., & Zhao, J. Y. (2001a). Biomarker identification by feature wrappers. *Genome Research, 11*(11), 1878–1887.

Xiong, M. M., Li, W. J., Zhao, J. Y., Li, J., & Boerwinkle, E. (2001b). Feature (gene) selection in gene expression-based tumor classification. *Molecular Genetics and Metabolism, 73*, 239–247.

Yang, K., Cai, Z., Li, J., & Lin, G. (2006). A stable gene selection in microarray data analysis. *BMC Bioinformatics, 7*, 228.

Yeo, G., & Poggio, T. (2001). Multiclass classification of SRBCTs, AI Memo 2001-018, CBCL Memo206.

Zainuddin, Z., & Ong, P. (2011). Reliable multiclass cancer classification of microarray gene expression profiles using an improved wavelet neural network. *Expert Systems with Applications, 38*, 13711–13722.

Zhang, M. L., & Zhou, J. H. (2007). ML-KNN: A lazy learning approach to multi label learning. *Pattern Recognition, 40*(7), 2038–2048.

Zheng, C. H., Huang, D. S., Zhang, L., & Kong, X. Z. (2009). Tumor clustering using nonnegative matrix factorization with gene selection. *IEEE Transactions on Information Technology in Biomedicine, 13*(4), 599–607.