Journal of Network and Innovative Computing
ISSN 2160-2174 Volume 4 (2016) pp. 236-245

DHS:An unsupervised feature selection algorithm based on Harmony Search for Microarray data Classification

K.Umamaheswari[1], M.Dhivya[2]

[1]Department of Information Technology,PSG College of Technology,Coimbatore, India
umakpg@gmail.com

[2]Department of Information Technology,SRM University,Chennai, India
dhivyapsg12@gmail.com

***Abstract*.Feature Selection is an essential task in microarray data classification.Various methods are available to handle the data with class labels whereas some data are mislabeled and unreliable. Unsupervised gene selection methods are existing to handle such data. We propose an unsupervised filter based method known as dynamic Harmony Search(DHS) which integrates Harmony Search into filter approach by defining new fitness function and it is independent of any learning model.The main aim of the filter approach is to quantify the relevance based on the intrinsic properties of the data.The proposed method is applied on benchmark microarray datasets and the results are compared with well known unsupervised gene selection methods using different classifiers. The proposed method governs promising enhancement on feature selection and good classification accuracy**.

Keywords:Microarray data, Feature reduction, Harmony Search

I. Introduction

Microarray data is widely and successfully applied to cancer classification of biomedical research. A typical microarray data set contains a large number of genes(tens of thousands to hundreds of thousands) and relatively few samples (less than one hundred).In the tens of thousands of genes, only a small part of the genes that help cancer classification. The main goal is to examine the differentially expressed genes as normal or abnormal. Doctors can make use of this information for disease diagnosis and treatment of patients. The most important challenge is "curse of dimensionality"[1] to be considered where the high dimensional nature of microarray data with less number of samples need to be analyzed .Major demerits of this nature are irrelevant and redundant genes. Data preprocessing techniques are used to obtain accurate and relevant information of genes[2,3].One of the most common data preprocessing techniques is feature selection which selects subset of genes from the original dataset and increases the performance.

Feature selection methods are categorized into four categories including filter, wrapper, embedded and hybrid approaches[3-6].The filter approaches are independent of learning model and estimate the relevance of genes based on the statistical properties of the data.There are various strategies available to assess the relevance of genes including univariate and multivariate strategies[2,6-8]. The univariate strategy assess and ranks the genes individually using a given criterion.Then the top ranked genes subset is chosen as the final subset. These methods are fast and efficient,sometimes the classification results are less accurate because it is unaware of correlation between genes. Some of the univariate methods are term variance [1], Laplacian score [9, 10], Signal-to-Noise ratio [11], mutual information [12] and information gain [13]. The multivariate strategy overcomes the drawback of univariate strategy by considering the correlation between genes. There are various multivariate strategies such as mRMR [14], FCBF [15], RSM [16], and Mutual correlation [17]. These strategies are single track search and results into the local optimum solutions.

The wrapper approach relies on a specific learning model in the gene selection process to assess a subset of selected genes and its accuracy is used to guide the search process. It falls into two approaches including greedy and stochastic search strategy [6]. The greedy search strategy is a single-track search and leads to local optimum results. Sequential forward selection and sequential backward selection are two basic methods used in the greedy search strategy [18]. The other method, the stochastic search strategy employs the randomness nature in the gene selection process. It includes ant colony optimization (ACO) [19], particle swarm optimization (PSO) [20] and genetic algorithm (GA) [21]. The performance of the wrapper approach is better than that of the filter approach, but this approach is subject to high computational cost especially for high-dimensionality of microarray datasets.

The third method is the embedded approach where a specific learning model is trained using an initial gene set to build a criterion, for ranking the values of genes. Some of the embedded based methods include support vector machine based on recursive feature elimination (SVM-RFE) [22] and random forest [23]. The main advantage of the embedded approach is the interaction with the learning model, but the computational time for training a specific classifier with the original gene set is high.

The hybrid approach evolves to combine the advantages of both filter and wrapper approaches. Initially subset of genes is chosen using the filter approach and the final gene set is selected based on the wrapper approach. Examples of the

hybrid approach are chi-square statistics with GA [24], information gain with memetic algorithm [25] and multiple-filter-multiple- wrapper (MFMW) method [26]. The major drawback of the hybrid approach is that the filter and the wrapper approaches are not seamless with each other and lead to degradation in classification performance.

Swarm intelligence-based methods such as ACO and PSO are multi-agent systems with the collective behavior of a population of artificial agents. These methods are considered to be very effective in selecting feature subset and have been successfully employed for the applications like face recognition [27], text classification [28] and financial domains [29]. A recent meta-heuristic technique Harmony Search works with solution vectors and shows significant results in feature gene selection[30].In microarray datasets, the wrapper approach is not mostly used due to time consumption. Therefore, the filter approach is suggested for the microarray data classification problem. If the class labels of the microarray data are available, then it is termed as supervised gene selection methods [11-14]. Neverthless, some of the microarray data samples incorrectly labeled or may have unreliable class labels [3, 31]. On that account, the significance of the unsupervised gene selection methods have been employed in the DNA microarray field.

The main objective of the proposed method is to build a system to combine the computational efficiency of the filter approach and the good performance of the Harmony Search(HS) algorithm, in which the learning model and the class labels of the sample are not needed in the gene selection process. In this paper, we propose a novel unsupervised filter based gene selection method for microarray data classification called microarray gene selection based on HS with dynamic genetic operators.Moreover, the performance of the chosen subsets of genes are evaluated using a new proposed fitness function independent of any learning model. Finally, the best subset of genes in all iterations are chosen as the final gene set.

The rest of the paper is composed as follows. Section 2 briefly reviews the harmony search algorithm. Section 3 describes the system design and Section 4 presents the proposed gene selection method using the HS algorithm. Section 5 and 6 provides the performance measures and experimental results on five microarray datasets respectively.

II.Related Work

The Harmony Search (HS) [32] is a meta-heuristic algorithm inspired by musical process of searching for a perfect shape of harmony. The algorithm is based on natural musical processes in which a musician searches for a better state of harmony by tuning pitch of each musical instrument, such as jazz improvisation. The music improvisation by pitch adjustment in the Harmony Search is analogous to local and global search process to find better solution in any optimization techniques.

The harmony memory (HM) is a group of pre-defined number of solution vectors similar to a population of particle in PSO or chromosome in GA. Initially HM is initialized with random solution vectors and the solution vectors in HM are improved using harmony search procedure known as HM improvisation step. This step is entirely controlled by the parameters: Harmony Memory Consideration Rate (HMCR), Pitch Adjustment Rate (PAR) and Bandwidth (bw).

In HS, the HMCR controls the balance between exploration and exploitation and it is set between 0 and 1. The searching procedure behaves as purely random search, if the HMCR is set to 0 and a value 1 for HMCR specifies 100% of previous solution vectors from HM are taken into consideration for next generation, which means, there is no chance to improve the harmony from outside the HM. In this way, HMCR keeps the balance between exploration and exploitation. Then the parameter PAR determines the rate of adjustment of solution vectors based on the bandwidth (bw) which is a variable and behaves as step size.

The HMCR and PAR determine Memory Consideration Probability (MCP), Pitch Adjustment Probability (PAP) and Random Probability (RP) as follows:

$$\mathrm{MCP} = \mathrm{HMCR} * (1 - \mathrm{PAR}) * 100 \quad (1)$$

$$\mathrm{PAP} = \mathrm{HMCR} * \mathrm{PAR} * 100 \quad (2)$$

$$\mathrm{RP} = 100 - \mathrm{MCP} - \mathrm{PAP} \quad (3)$$

Basically, improvisation of HM is governed by these parameters (MCP, PAP, and RP).In HS, the bw and PAR are fixed and pitch adjustment is done according to Eq. (4).

$$\mathrm{HM_i}(t+1) = \begin{cases} \mathrm{HM_i}(t+1) = \mathrm{HM_j}(t) - \mathrm{rand}(1) * \mathrm{bw\ if\ rand}(1) < 0.5 \\ \mathrm{HM_i}(t+1) = \mathrm{HM_j}(t) + \mathrm{rand}(1) * \mathrm{bw\ if\ rand}(1) > 0.5 \end{cases} \quad (4)$$

In Eq.(4), $\mathrm{HM_i}(t+1)$ is the next ith harmony at time t + 1 and $\mathrm{HM_j}(t)$ is the jth randomly selected harmony for pitch adjustment at time t.

HS algorithm has been very successful in a wide variety of optimisation problems, presenting several advantages with respect to traditional optimisation techniques. It imposes only limited mathematical requirements and is not sensitive to the initial value settings. The HS algorithm generates a new potential solution vector, after considering all existing vectors. This technique has been applied to damper location in structural system[33], electrical engineering [34,35], economics[36], transport[37,38], ecology[39,40],biomedical [41] and pipe design problem[42].Geem[43] applied multi-objective optimization for the design of a satellite heat pipe.

HS method is a random search technique. It does not require any prior domain information, such as the gradient of the objective functions. It is different from other population-based evolutionary approaches, it only utilizes a single search memory to evolve. Therefore, the HS method has the characteristics of algorithm simplicity, convergence speed and easy implementation. On the other hand, it has a drawback like weak local search ability.To overcome, dynamic genetic operators are combined with harmony search algorithm.

III.System Design

In this paper,benchmark Microarray datasets are used for gene selection using unsupervised filter based HS.The framework of the proposed gene selection process is shown in Fig.1. The proposed method is a population based metaheuristic algorithm initialized with size of harmony memory, HMCR, PAR,number of iterations and solution vectors. Fitness function is determined for all solution vectors which is independent of any learning model.A new harmony vector is

generated based on memory consideration,pitch adjustment and random selection.

The process of generating new harmony vector is called improvisation. Every component obtained by the memory consideration is examined to determine whether it should be pitch-adjusted. This operation uses the PAR parameter, which is the rate of pitch adjustment. The pitch adjustment or random selection is applied to each variable of the new harmony vector. If the new harmony vector is better than the worst harmony in the HM, judged in terms of the fitness value, the new harmony is included in the HM and the existing worst harmony is excluded from the HM.Then the genetic operators like crossover and mutation are used dynamically in order to reduce the execution time and select best solutions.The above process is iterated till either one of the termination criteria is reached (1) harmony achieves 100% classification accuracy or (2) number of iterations greater than 100. The best particles are estimated on different types of classifiers including SVM,Naïve Bayes and Decision Tree.

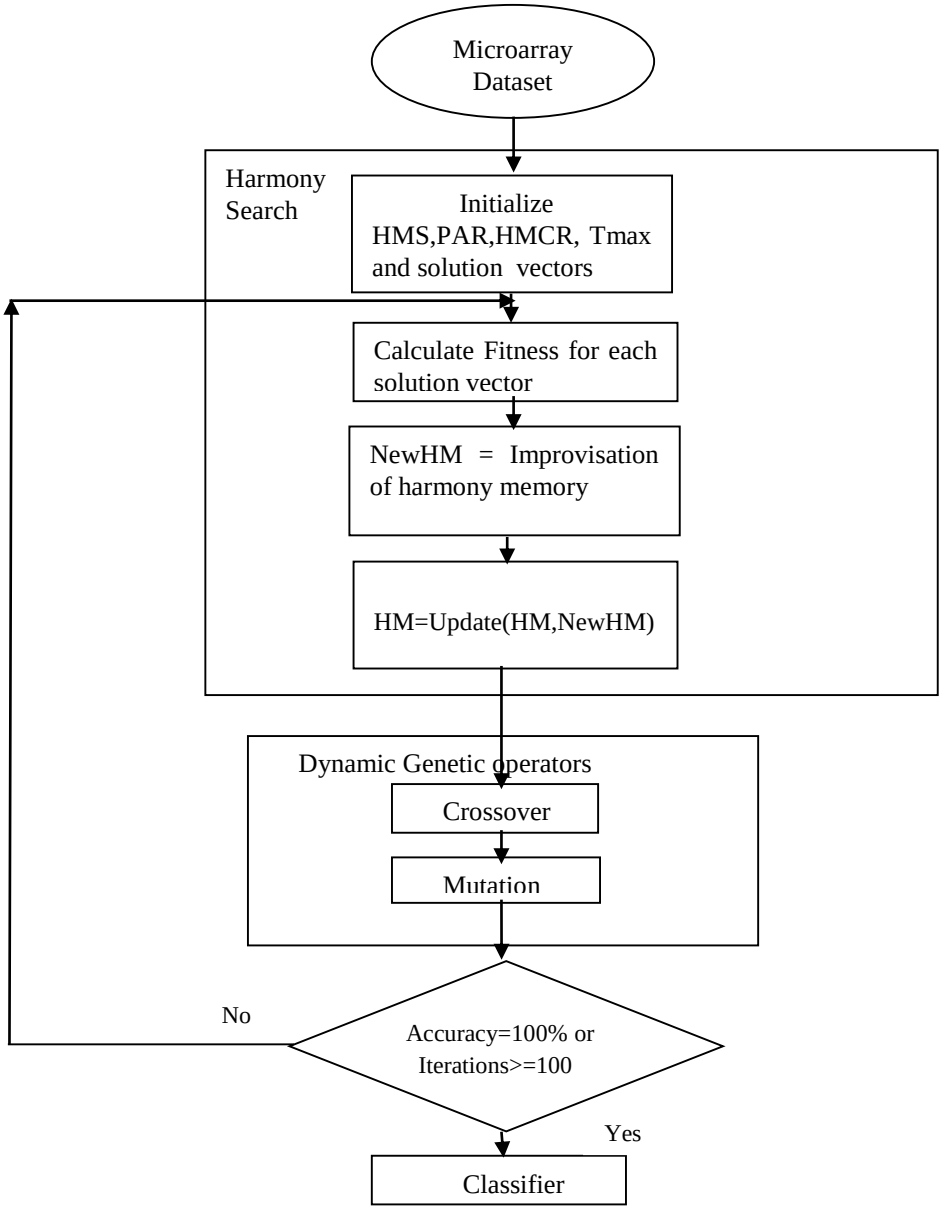


Figure 1. Framework of the proposed algorithm

IV.Proposed Method

The proposed method includes two parts:feature selection and classification.Unsupervised Filter based Feature Selection is done using the integrated technique of HS for enhancement of solutions and genetic dynamic operators for reproducing best solutions.The fitness function is independent of any learning model so it shows good performance over different type of classifiers including SVM,Naïve bayes and Decision Tree.The following steps clearly elaborate the proposed method:

Proposed Algorithm

Begin

1.Initialize algorithm parameters size of harmony memory,HMCR,PAR
2.Preprocess and split data into training and test
3.Generate initial harmony randomly
4.while(termination criteria not met)do
5. Evaluate the fitness of all the solution vectors
6. for i=1 to number of solution vectors do
7. create new harmony by adjusting parameters
8. if new harmony vector is better than the worst harmony in the HM
9. include the new harmony in HM and exclude existing worst harmony
10. update harmony memory
11. end if
12. end for
13. Calculate dynamic parameter crossover rate and mutation rate
14. end while
15. Calculate the classification accuracy rate

1.Encoding and Initialization: The features are encoded as binary values (0 or 1) where '1' represents the feature to be selected and '0' represents the feature not selected. The parameters like harmony memory size, HMCR and PAR are initialized.Initial harmony will be taken as the randomly filled solution vectors in harmony memory and then it will be updated based on the fitness value.

2.Fitness Function:The filter based method is to quantify the relevance between the intrinsic properties of the genes. Thereafter, subsets with maximum relevance should get a greater fitness value. The fitness value of solution k is computed as follows:

$$fitness(k) = \frac{1}{|subset(k)|}\sum_{i=1}^{|subset(k)|} relevance\left(g_i^k\right) \quad (5)$$

where subset(k) is the subset of genes selected by solution vector k, $|subset(k)|$is the size of subset(k), g_i^k is the ith gene in the subset(k) and relavance is the function that evaluates the relevance of each gene. In this paper the term variance [1] is used as a relevance function, which is defined as follows:

$$TV(g_i) = \frac{1}{p}\sum_{s=1}^{p}(g_{is} - \bar{g}_i)^2 \quad (6)$$

where p is the number of samples, g_{is} denotes the value of gene i for sample s, and $\bar{g}_i$is the average value of all the samples corresponding to gene g_i. Also, the relevance value of each gene is normalized in the interval [0..1] using the softmax scaling function [1]. Note that the number of selected genes by solution vectors in each iteration is equal to a constant value NG. It can be seen from Eq. (5) that this specific kind of fitness function is independent of any learning model.

3.Updation: For each solution vector, its new harmony fitness value is compared with the fitness of existing harmony, if the new harmony value performs better than the worst one in the HM, the new one is included in the HM and the corresponding worst candidate is excluded.

4. Dynamic crossover and mutation rate. The values of the crossover and mutation rates are set dynamically. The crossover rate for two chromosomes is determined by the fitness values of the two chromosomes. The mutation rate of a chromosome is calculated only by the fitness value of the chromosome. The formulae for the crossover and mutation rates are shown as follows:

$$D_c = \begin{cases} \alpha\left[1 - \frac{f - f_{med}}{f_{max} - f_{med}}\right] & for f > f_{med}, \\ \alpha & for\ f \le f_{med}; \end{cases} \quad (7)$$

$$D_m = \begin{cases} \beta\left[1 - \frac{f_{mut} - f_{med}}{f_{max} - f_{med}}\right] & for f_{mut} > f_{med}, \\ \beta & for\ f_{mut} \le f_{med}; \end{cases} \quad (8)$$

Where
D_c denotes the crossover rate
D_m denotes the mutation rate
f denotes the largest fitness value of the two chromosomes in a crossover operation
f_{mut} denotes the fitness value of the chromosome in a mutation operation
f_{med} and f_{max} are the median and maximum fitness values

The values of both α and β are set to 1.According to Eq.(7), for a pair of chromosomes with small fitness value, high crossover rate is assigned to increase their chance of evolution. When the highest fitness values of a pair of chromosomes is less than or equal to the median fitness value of the current population, crossover rate of 1 is assigned to make them to evolve. Similarly, from Eq.(8), higher mutation rate is assigned for a chromosome with lower fitness value.

5.Termination Criteria: The steps 2-4 are iterated when any one of these two termination conditions statisfied: (1)when the fitness value of one harmony in the current generation achieves 100% classification accuracy, or (2) when the number

of iterations is larger than 100 or the best fitness value of the last 15 iterations remains the same.

6.Classification:The final gene subset is classified using different types of classifiers including SVM,Naïve Bayes and Decision Tree.

V.Performance Measures

Initially, the features of datasets used to have more variations and deviations. They also consist of missing values which are resolved using data pre-processing techniques with the values normalized. Harmony size is defined and solution vectors are generated randomly.Harmony Search parameters are initialized and then fitness function is calculated.The result of change of fitness value is considered and assign its harmony.Dynamic crossover and mutation rate is applied. Once the optimal best features are obtained, the classification accuracy is evaluated using test set.

While classifying the data, the obtained outcomes are based on the confusion matrix which consists of the terms like True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). Based on the defined terms, there are some classification performance metrics available namely Accuracy, Precision, Specificity, Sensitivity, t-test and F-measure. Based on the above metrics, accuracy is the most widely used metric to analyse the classification problem and it is calculated using:

$$Classifier\ Accuracy = \frac{TP + TN}{No.of\ Positives + No.of\ Negatives} \quad (9)$$

VI.Experimental Results

A.Dataset Description

The datasets collected for our experiments are Leukemia[44], Lung[45], Colon[44], Prostrate[45] and SRBCT[45]. Table 1 lists the details of these data sets , including the sample size, the number of gene expression variables and number of classes with description for each data set .

Table 1. Dataset Description

Dataset	No. of Samples	No. of genes	No. of Classes	Description
Leukemia	72	7129	2	47 ALL , 25 AML
Lung	203	12600	5	139 ADCA, 21 Squamous cell, 20 Pulmonary carcinoids, 6 SCLC, 17 normal
Colon	62	2000	2	40 Tumor, 22 Normal
SRBCT	83	2308	4	29 EWS,11 BL, 18 NB, 25 RMS
Prostrate	102	10509	2	52 tumor,50 normal

B.Results and discussion

The experiment is done using Java with NetBeans IDE 7.1 in Windows 7.The analysis results of various stages are presented below:

Assumptions:

- The size of harmony memory is 100
- Maximum number of iterations are 50
- The value of HMCR,PAR and bandwidth are set to 0.9,0.3 and 0.0001 respectively.

The WEKA machine learning software library [46] is used for the implementation of the classifiers. SMO with the polykernel is chosen as the SVM classifier which used the one-against-rest strategy for the multiclass problems. The complexity parameter c is set to 1 and the tolerance parameter is set to 0.001. Moreover, naïve bayes is used as the NB classifier. Furthermore, J48 is adopted as the DT classifier, in which the post-pruning technique is used in the pruning phase where its confidence factor is set to 0.25 and the minimum number of samples per leaf is set to 2.

Table 2. Average Classification accuracy of datasets over 5 independent runs using SVM classifier

Datasets	Avg No. of selected genes	Classification Accuracy (%)						
		DHS	D-MBPSO	RSM	MC	RRFS	TV	LS
Colon	16	78.29	77.09	75.46	61.82	75.46	78.19	66.37
SRBCT	19	76.53	79.32	62.07	54.49	68.28	60.69	63.45
Leukemia	23	65.11	60.04	62.36	61.77	76.48	79.41	64.71
Prostrate	28	63.62	61.78	77.15	65.72	69.15	72	52
Lung	14	87.47	85.34	64.29	71.43	80.86	72.29	82
Average		**74.20**	**72.71**	**68.27**	**63.05**	**74.05**	**72.52**	**65.71**

Table 3. Average Classification accuracy of datasets over 5 independent runs using Naïve Bayes classifier

Datasets	Avg No. of selected genes	Classification Accuracy (%)						
		DHS	D-MBPSO	RSM	MC	RRFS	TV	LS
Colon	16	85.33	87.29	73.64	68.19	67.28	58.19	52.73
SRBCT	19	88.45	89.12	62.08	62.07	71.73	61.38	67.59
Leukemia	23	64.87	60.04	57.65	70.59	64.71	67.65	91.18
Prostrate	28	77.26	75.07	69.72	66.29	68.58	66.86	67.43
Lung	14	68.12	66.83	76.43	40.96	78.29	68.01	70.01
Average		**76.81**	**75.67**	**67.90**	**61.62**	**70.12**	**64.42**	**69.79**

Table 4. Average Classification accuracy of datasets over 5 independent runs using Decision Tree classifier

Datasets	Avg No. of selected genes	Classification Accuracy (%)						
		DHS	D-MBPSO	RSM	MC	RRFS	TV	LS
Colon	16	76.44	78.65	71.82	66.37	65.46	68.19	60.91
SRBCT	19	70.67	73.43	41.38	55.87	71.04	77.25	54.49
Leukemia	23	73.14	72.31	61.18	67.65	79.42	79.42	70.59
Prostrate	28	65.58	68.96	66.29	64	62.29	61.15	56.01
Lung	14	83.22	84.35	69.29	68.58	79.72	75.72	78.57
Average		**73.81**	**75.54**	**61.99**	**64.49**	**71.59**	**72.35**	**64.11**

The performance of the proposed method is evaluated over various datasets with different types of classifiers. Table 2-4 represents the comparisions of proposed method over various unsupervised filter based method like D-MBPSO[47],RSM, MC,RRFS,TVand LS .The average classification accuracy of the datasets over 5 independent runs using SVM,naïve bayes and decision tree algorithms and evaluated. The average classification accuracy are shown in the last row of the table.

It is inferred from Table 2 that the proposed method obtains the highest classification accuracy of 78.29% for colon,65.11% for Leukemia,63.62% for Prostrate,87.47% for Lung whereas for SRBCT the accuracy obtained 76.53% less than D-MBPSO.The average classification accuracy over all of the datasets show that the proposed method with an accuracy of 74.20% outperforms D-MBPSO ,RSM ,MC ,TV and LS ,where RRFS shows 0.15% increase of accuracy than the proposed method.

The results of Table 3 illustrate that proposed method outperforms the other methods in terms of classification accuracy for the NB classifier on the Colon, SRBCT, Leukemia, and Lung Cancer datasets. The average values on all of the datasets, in the last row of Table 3, show that the DHS is superior to all the other methods. It outperforms D-MBPSO by 1.14%, RSM by 8.91%, MC by 15.19%, RRFS by 6.69%, TV by 12.39%, and LS by 7.02%.

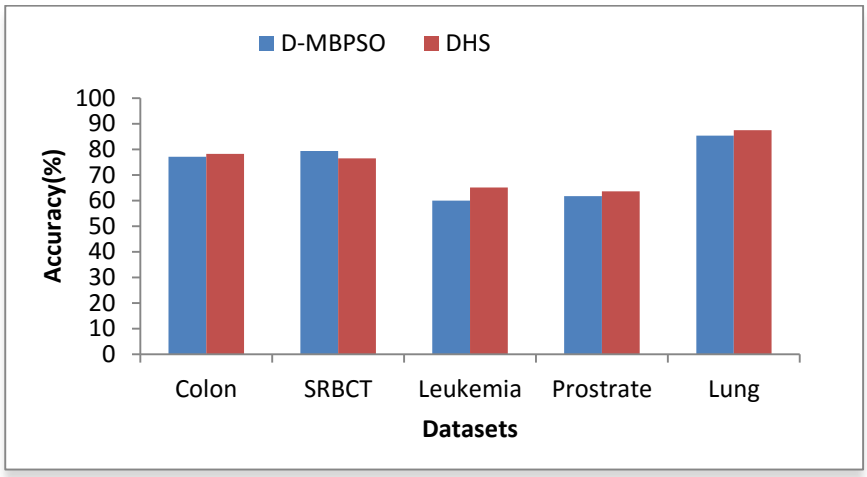
From Table 4 it is observed that the classification accuracy of the DT classifier based on the proposed method is superior to that of the unsupervised filter-based methods as 73.14% for Leukemia but 76.44% for Colon, 65.58% for Prostate Tumor ,70.67% for SRBCT and 83.22% for Lung Cancer datasets where it gives statisfied results than proposed method. The average values in the last row of Table 4 show that the proposed method outperforms RSM by 11.82%, MC by 9.32%, RRFS by 2.22%, TV by 1.46%, and LS by 9.7%, where D-MBPSO shows 1.73% increase of accuracy than the proposed method.

It can be concluded from Tables 2-4 that proposed method shows an improvement of 5-6% over the existing unsupervised filter-based methods (i.e., D-MBPSO, RSM, MC, RRFS, TV, and LS) in terms of classification accuracy for each of the three classifiers over different datasets.

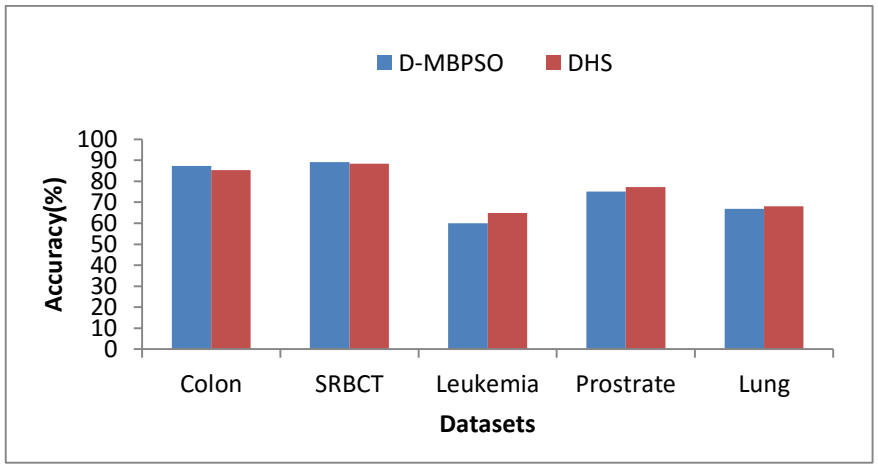
The performance of the proposed method has been evaluated over different numbers of selected genes using various types of classifiers. Figures 2(a)-(c) report the graphical results of the different datasets using the SVM, NB, and DT classifiers correspondingly. The x-axis denotes the type of datasets, while the y-axis shows the average classification accuracy (in %).

Fig. 2(a) shows the results of proposed method with D-MBPSO using SVM classifier which gives desired results for lung,leukemia and prostrate datasets. It can be concluded that the classification error rate of the proposed method is significantly superior to that of the D-MBPSO method for colon,lung,prostrate, leukemia and less significant results for SRBCT dataset. As seen in Fig. 2(b) illustrates the respective comparison results for different datasets with proposed method and D-MBPSO using naïve bayes classifier. The different classification accuracy rates of proposed method and D-MBPSO can be seen more prominently for SRBCT and Colon dataset where the proposed method acquires significantly higher classification accuracy for Leukemia, Prostrate and Lung as 64.87%,77.26% and 68.12%, correspondingly, than D-MBPSO.

2(a)



(b)



(c)

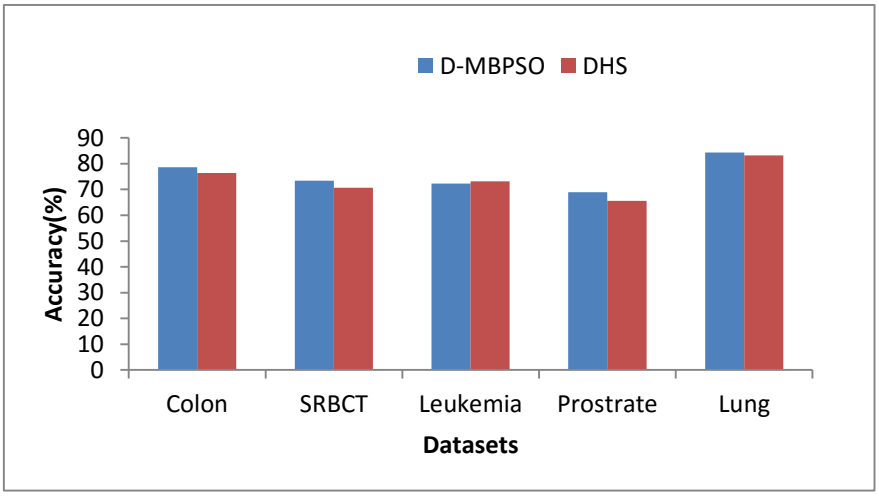


Figure 2. Classification accuracy for five datasets using classifiers: (a) SVM (b) NB (c) DT

Fig. 2(c) demonstrates that the overall performance of proposed method is superior to that of the D-MBPSO method when the decision tree classifier is applied. Especially, for Colon,SRBCT,Prostrate and Lung datasets, the classification accuracy of the proposed method is 76.44%,70.67%,65.58% and 83.22%, while for the D-MBPSO this value is reported

as 78.65%,73.43%,68.96% and 84.35% respectively. Moreover, it can be seen that the performance of the proposed method is 0.83% better than that of D-MBPSO for Leukemia and the proposed method obtains less significant results than D-MBPSO for other datasets.

It can be concluded from Figs. 2(a)-(c) that although the proposed method is an unsupervised method and does not need class labels of the samples, it can be much better than the D-MBPSO method.The proposed method is a population based method which simultaneously explores the search space from different points.

VII.Satistical Analysis

In order to illustrate that the experimental results are statistically significant, the Friedman test [48] has been performed on the results. The Friedman test is a non-parametric test used to measure the statistical differences of methods over multiple datasets. For each dataset, the methods are ranked separately based on the classification accuracy.The method with the highest classification accuracy gets rank 1, the second lowest result gets rank 2, and so on. When several methods have the same classification accuracy, their average rank is assigned to each method. The Friedman test is distributed according to the Fisher distribution with (k-1)and (k-1)(N-1) degrees of freedom which is defined as follows:

$$F_F = \frac{(N-1)\chi_F^2}{N(k-1)-\chi_F^2} \quad (13)$$

Where

$$\chi_F^2 = \frac{12N}{K(k+1)}\left[\sum_{j=1}^{k} R_j^2 - \frac{k(k+1)^2}{4}\right] \quad (14)$$

N is the number of datasets, k is the number of methods, and Rj is the average rank of the j-th method over all datasets. The null hypothesis in Friedman test means that all methods perform equally at the significance level α. The null hypothesis is accepted when FF is less than the critical value; otherwise it is rejected. In the experiments, the significance level was set to α=0.05. The average ranks of the unsupervised filter-based methods using SVM, NB, and DT classifiers are calculated according to the values in Tables 2-4.

Table 5. The results of Friedman test for SVM,NB and DT classifiers

Classifier	χ_F^2	F_F	F(6,24)	Significant
SVM	8.785	1.656	2.51	=
NB	16.528	4.907	2.51	+
DT	11.971	2.655	2.51	+

In the experiments, N=5 ,k=7 and the critical value of Fisher distribution (7-1)=6 and (7-1) (5-1) =24 with degrees of freedom is equals to F(6,24)=2.51 . From Table 5,it is inferred that the gene selection methods are integrated with NB and DT classifiers, the value of FF is greater than 2.51. Therefore, the null hypothesis will be rejected and it can be concluded that these results are statistically significant. Moreover, the value of FF is less than 2.51 when SVM classifier is used. Therefore, the null hypothesis is accepted and it is clear that proposed method performs equally with the other unsupervised filter based methods.

VIII.Conclusion

In this paper,unsupervised filter based method is proposed known as dynamic HS based on Harmony Search mechanism for gene selection process. The computational efficiency of the filter approach and the HS are combined to improve the performance of the proposed method. Moreover, a new fitness function is used to evaluate the subsets of selected genes without using any learning model to enhance the efficiency of the proposed method. The performance of the proposed method is examined on the five microarray datasets using three different classifiers including support vector machine, naïve Bayes, and decision tree. Then, the proposed method is also compared to the well known unsupervised filter-based gene selection methods including unsupervised feature selection based on PSO (D-MBPSO), relevance-redundancy feature selection (RRFS), random subspace method (RSM), mutual correlation (MC), term variance (TV), and Laplacian score (LS). It is found that the classification accuracy of the proposed method outperforms other unsupervised methods for various subsets of genes over all the three classifiers. The results obtained using the proposed method is significantly better which is proven using F test and could be used in the prediction analysis of medical application field.

References

[1] S. Theodoridis, K. Koutroumbas, *Pattern Recognition*, fourth ed., Elsevier Science, 2008.

[2] C. Lazar, J. Taminau, S. Meganck, D. Steenhoff, A. Coletta, C.Molter, V.d.Schaetzen,R. Duque, H.Bersini, A.Nowe,“ A Survey on Filter Techniques for Feature Selection in Gene Expression Microarray Analysis”, *IEEE/ACM Transaction on Computational Biology and Bioinformatics*, 9 ,1106-1119,2012.

[3] V.Bolón-Canedo, N. Sánchez-Maroño, A.Alonso-Betanzos, J.M.Benítez,F. Herrera,“A review of microarray datasets and applied feature selection methods*”*, *Information Sciences*, 282 ,2014.

[4] Hochreiter S, Obermayer K., “ Kernel Methods in Computational Biology”.*In Scholkopf* B, Tsuda K, Vert JP, editors. MIT press, pp. 323,2004.

[5] M. Dash ,H. Liu,“Feature Selection for Classification”, *Intelligent Data Analysis: An Int'l J.*, 1(3),131- 156,1997.

[6] Y. Saeys, I. Inza, P. Larrañaga, “ A review of feature selection techniques in bioinformatics “, *Bioinformatics*, 23, 2507-2517,2007.

[7] S. Tabakhi, P. Moradi, F. Akhlaghian,“An Unsupervised feature selection algorithm based on ant colony optimization”, *Engineering Applications of Artificial Intelligence*, 32 ,112-123,2014.

[8] C. Lai, M. Reinders, L. van't Veer, L. Wessels, “A comparison of univariate and multivariate gene selection techniques for classification of cancer datasets”, *BMC Bioinformatics*, 7 ,235,2006.

[9] B. Liao, Y. Jiang, W. Liang, W. Zhu, L. Cai, Z.Cao, “Gene selection using locality sensitive Laplacian score” , *IEEE/ACM Transactions on Computational*

Biology and Bioinformatics, 1146-1156, 2014.

[10] X. He, D. Cai, P. Niyogi, " Laplacian Score for Feature Selection ", *Advances in Neural Information Processing Systems*, 18,507-514,2005.

[11] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M.Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D.Bloomfield, E.S. Lander, "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring ", *Science*, 286,531- 537, 1999.

[12] R. Cai, Z. Hao, X. Yang, W. Wen, " An efficient gene selection algorithm based on mutual information", *Neurocomputing*, 72 ,991-999,2009.

[13] L.E. Raileanu, K. Stoffel, "Theoretical Comparison between the Gini Index and Information Gain Criteria" , *Annals of Mathematics and Artificial Intelligenc*e, 41, 77-93,2004.

[14] C.Ding,H. Peng, "Minimum redundancy feature selection from microarray gene expression data", *Journal of Bioinformatics and Computational Biology*, 03,185-205,2005.

[15] L.Yu,H.Liu,"Feature Selection for High-Dimensiona Data: A Fast Correlation-Based Filter Solution", *in: 20th International Conference on Machine Learning*, pp. 856-863,2003.

[16] C.Lai,M.J.T. Reinders,L.Wessels,"Random Subspace method for multivariate feature selection", *Pattern Recognition Letters*, 27 ,1067-1076,2006.

[17] M.Haindl,P.Somol,D.Ververidis, C. Kotropoulos, " Feature Selection Based on Mutual Correlation" , in: *Pattern Recognition,Image Analysis and Applications*,Springer Berlin Heidelberg, pp. 569-577,2006.

[18] I. Inza, B. Sierra, R. Blanco, P. Larra ñaga, "Gene selection by sequential search wrapper approaches in microarray cancer class prediction" , *Journal of Intelligent and Fuzzy Systems*, 12 ,25-33,2002.

[19] Y. Li, G.Wang, H. Chen, L. Shi, L.Qin, "An Ant Colony Optimization Based Dimension Reduction Method for High-Dimensional Datasets", *Journal of Bionic Engineering*, 10 ,231-241,2013.

[20] B. Sahu, D. Mishra, "A Novel Feature Selection Algorithm using Particle Swarm Optimization for Cancer Microarray Data", *Procedia Engineering*, 38, 27-31,2012.

[21] C.H. Ooi, P. Tan, "Genetic algorithms applied to multi-class prediction for the analysis of gene expression data", *Bioinformatics*, 19 ,37-44,2003.

[22] I.Guyon, J. Weston, S. Barnhill, V. Vapnik, " Gene Selection for Cancer Classification using Support Vector Machines", *Machine Learning*, 46, 389-422,2002.

[23] R. Diaz-Uriarte, S. Alvarez de Andres, "Gene selection and classification of microarray data using random forest", *BMC Bioinformatics*, 7 ,3,2006.

[24] C.-P. Lee, Y. Leu, "A novel hybrid feature selection method for microarray data analysis", *Applied Soft Computing*, 11 ,208-213,2011.

[25] A. Zibakhsh, M.S.Abadeh,"Gene selection for cancer tumor detection using a novel memetic algorithm with a multi-view fitness function ", *Engineering Applications of Artificial Intelligence*,26,1274- 1281, 2013.

[26] Y. Leung, Y. Hung, "A Multiple-Filter-Multiple-Wrapper Approach to Gene Selection and Microarray Data Classification" , *IEEE/ACM Transaction on Computational Biology and Bioinformatics*,7,108-117, 2010.

[27] H.R.Kanan,K.Faez,"An improved feature selection method based on ant colony optimization (ACO) evaluated on face recognition system", *Applied Mathematics and Computation*, 205,716-725,2008.

[28] M.H.Aghdam,N.Ghasem-Aghaee,M.E.Basiri,"Text feature selection using ant colony optimization", *Expert Systems with Applications*, 36,6843-6853,2009.

[29] Y.Marinakis,M. Marinaki, M. Doumpos, C. Zopounidis, " Ant colony and particle swarm optimization for financial classification problems", *Expert Systems with Applications*, 36 ,10604-10611,2009.

[30] Jun Wei, "Novel global harmony search algorithm for feature gene selection" , *Journal of Chemical and Pharmaceutical Research*, 7(3),2201-2207,2015.

[31] S. Niijima, Y. Okuno, "Laplacian Linear Discriminant Analysis Approach to Unsupervised Feature Selection",*IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 6 , 605- 614,2009.

[32] Geem ZW, Kim JH, Loganathan GV, "A new heuristic optimization algorithm: harmony search", *Simulation*, 76(2),60–68,2001.

[33] Fereidoun Amini, Pedram Ghaderi," Hybridization of Harmony Search and Ant Colony Optimization for optimal locating of structural dampers", *Applied Soft Computing*, 13 ,2272–2280,2013.

[34] X. Z. Gao, V. Govindasamy, H. Xu, X. Wang, and K. Zenger, "Harmony Search Method: Theory and Applications " , *Computational Intelligence and Neuroscience*, vol. 2015, Article ID 258491, 10 pages, 2015. doi:10.1155/2015/258491

[35] Rong, Z. and L. Hanzo, "Iterative Multiuser Detection and Channel Decoding for DS-CDMA Using Harmony Search",*Signal Processing Letters, IEEE*,16(10), 917-920,2009.

[36] Jaberipour, M. and E. Khorram, "Solving the sum-of ratios problems by a harmony search algorithm", *Journal of Computational and Applied Mathematics*, 234(3),733-742,2010.

[37] Geem, Z.W., "Application of Harmony Search to Vehicle Routing" , *American Journal of Applied Sciences*, 1552-1557,2005.

[38] Mahdavi, M., " Solving NP-Complete Problems by Harmony Search", *in Music-Inspired Harmony Search Algorithm*, Springer Berlin / Heidelberg,53-70,2009.

[39] Geem, Z.W., Williams, J.C., "Harmony search and ecological optimization" , *International Journal of Energy and Environment*, 1, 150 – 154, 2007.

[40] Geem, Z., C.-L. Tseng, and J. Williams, "Harmony Search Algorithms for Water and Environmental Systems " , *in Music-Inspired Harmony Search Algorithm*, Springer Berlin / Heidelberg,113-127,2009.

[41] Mohsen, A., A. Khader, and D. Ramachandram, "An Optimization Algorithm Based on Harmony Search for

RNA Secondary Structure Prediction ", *in Recent Advances in Harmony Search Algorithm*, Springer Berlin /Heidelberg ,163-174,2010.

[42] Yoo Do Gluen,Lee Ho Min,Lee Eui Hoon,Kim Joong Hoon, " Efficiency Evaluation of Harmony Search Algorithm according to constraint handling techniques: Application to optimal pipe design problem ", *Journal of the Korea Academia-Industrial cooperation Society*,16(7) 4999-5008,2015.

[43] Zoon Woo Geem, "Artificial Satellite Heat Pipe Design Using Harmony Search ", *Advances in Intelligent Systems and Computing*,382,423-433,2015.

[44] Dataset Repository, Bioinformatics Research Group, Available:<http://www.upo.es/eps/bigs/datasets.html>, 2014.

[45] A . Statnikov, C.F. Aliferis, I. Tsamardinos, Gems: Gene Expression Model Selector, Available: < http://www.gems-system.org/>, 2005.

[46] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P . Reutemann, I. Witten, The WEKA data mining software, Available: <http:// www.cs. waikato .ac.nz/ml/weka>.

[47] K .Umamaheswari, M.Dhivya, "D-MBPSO: An Unsupervised Feature Selection Algorithm Based on PSO", *Innovations in Bio-Inspired Computing and Applications*, Advances in Intelligent Systems and Computing ,424,359-369,2015.

[48] M. Friedman, " The use of ranks to avoid the assumption of normality implicit in the analysis of variance" , *Journal of the American Statistical Association*, 32 ,675-701,1937.

Author Biographies

Dr. K. Umamaheswari received her Bachelor's degree in Computer Science and Engineering from Bharathidasan University in 1989 and her master's in Computer Science and Engineering from Bharathiar University in 2000. She received her PhD in Anna University- Chennai in 2010. She has rich experience in teaching for about 20 years and currently working as Professor in the department of Information Technology of PSG College of Technology, Coimbatore. Her research areas include Classification techniques in Data Mining and other areas of interest are Information Retrieval, Software Engineering, Theory of Computation and Compiler Design.

Dhivya. M received her Bachelor's degree in Information Technology from Anna University in 2009 and her Master's in Information Technology from Anna University in 2011. She is doing her Ph.D at PSG research center, Anna University, Chennai and working as Assistant Professor in the department of Information Technology, SRM University, Chennai. Her research interests are Data Mining,Computational Intelligence and Soft Computing.