Computing and Informatics, Vol. 31, 2012, 921–938

CLUSTERING IN CONJUNCTION WITH WRAPPER APPROACH TO SELECT DISCRIMINATORY GENES FOR MICROARRAY DATASET CLASSIFICATION

Rajni Bala

Deen Dayal Upadhyaya College
University of Delhi
New Delhi India
e-mail: r_dagar@yahoo.com

Ramesh Kumar Agrawal

School of Computer and Systems Sciences
Jawaharlal Nehru University
New Delhi India 110067
e-mail: rka@mail.jnu.ac.in

Communicated by Irena Roterman-Konieczna

Abstract. With the advent of microarray technology, it is possible to measure gene expression levels of thousands of genes simultaneously. This helps us diagnose and classify some particular cancers directly using DNA microarray. High-dimensionality and small sample size of microarray datasets has made the task of classification difficult. These datasets contain a large number of redundant and irrelevant genes. For efficient classification of samples there is a need of selecting a smaller set of relevant and non-redundant genes. In this paper, we have proposed a two stage algorithm for finding a set of discriminatory genes responsible for classification of high dimensional microarray datasets. In the first stage redundancy is reduced by grouping correlated genes into clusters and selecting a representative gene from each cluster. Maximal information compression index is used to measure similarity between genes. In the second stage a wrapper based forward feature selection method is used to obtain a set of discriminatory genes for a given classifier. We have investigated three different techniques for clustering and four classifiers in our experiments. The proposed algorithm is tested on six well known publicly available datasets. Comparison with the other state-of-the-art methods show that our proposed algorithm is able to achieve better classification accuracy with less number of genes.

Keywords: Cancer classification, microarray, clustering, representative entropy, maximal information compression index, gene selection

1 INTRODUCTION

With the advent of microarray technology, it is possible to measure gene expression levels of thousands of genes simultaneously. This has helped diagnose and classify some particular cancers directly using DNA microarray. In literature, a large number of classifiers have been employed for classification of cancerous vs non-cancerous patients or different types of cancers. Since microarray datasets are characterized by high dimension and small sample size classification is very difficult [1]. Many of these genes are irrelevant or redundant. Irrelevant genes not only increase size of the search space but also make generalization more difficult. Hence to build up an efficient classifier, we need to select a set of discriminatory genes by removing redundant, non-relevant and noisy genes. This process of selecting a set of discriminatory genes is known as gene selection and is carried out before building a classifier.

In the past a number of gene/feature selection methods have been suggested. These algorithms designed with different evaluation criteria broadly fall into two categories: the filter model and the wrapper model [2]. Most filter methods employ statistical characteristics of data for feature selection. These methods independently measure the importance of features without involving any learning algorithm; hence less computation is needed. Based on the evaluation criterion, filter methods are broadly categorized into two groups:

1. univariate and
2. multivariate evaluation methods.

Univariate methods consider the contribution of individual genes to the classification independently whereas multivariate methods consider the combined effect of genes for classification. Most of the methods suggested in literature are univariate methods. In univariate method, individual features are scored and ranked based on certain statistical criteria and features with highest ranking values are selected. In literature, it has been observed that the combination of individual good genes does not necessarily lead to good classification performance. Also, since genes selection methods do not consider the correlation among genes, gene subset so obtained may contain redundant genes. The wrapper model requires one predetermined learning algorithm and uses its performance as the evaluation criterion. It searches for features better suited to the learning algorithm aiming to improve classification performance. Wrapper methods are computationally more intensive than filter methods because they evaluate each candidate gene subset using a learning algorithm. The conventional wrapper methods have been applied for feature selection on small or medium scale datasets; but, due to large computation time, it is difficult to apply them directly on high dimensional datasets. Reducing the search space for wrapper methods will decrease the computation time. This can be achieved by selecting a smaller set of non-redundant features from the original set of features without losing any informative feature.

Recently some hybrid approaches [3, 4] have been suggested for gene selection in literature. They combine more than one approach to obtain a set of genes that provides better classification accuracy. In this paper, we have proposed a hybrid algorithm for selecting a set of discriminatory genes. In the first stage we reduce redundancy by grouping correlated genes. To achieve this, we propose to carry out gene clustering before gene selection. Various clustering techniques are suggested in literature. In this paper, we have investigated three different clustering techniques. Maximal information compression index is used for clustering correlated genes. From each cluster a representative gene is chosen to obtain a set of non-redundant genes. In the second stage the Sequential Forward feature selection method is applied to the set of genes obtained in the first stage for further selecting a smaller set of discriminatory genes. Finally, we evaluated the performance of these clustering algorithms in terms of classification accuracy, number of genes selected and computation time.

This paper is organized as follows: Section 2 describes briefly the clustering algorithms used. Section 3 presents outline of our proposed algorithm for gene selection. Experimental results on six well-known publicly available datasets are presented in Section 4. Section 5 contains conclusions.

2 GENE SELECTION USING CLUSTERING

Microarray datasets are characterized by large number of genes and very small size samples. Due to large dimension and small sample size, it suffers from curse of dimensionality [1]. Hence for building efficient classifier there is a need for selecting a smaller set of discriminatory genes. A variety of gene selection techniques have been proposed to determine relevant genes. The most commonly used gene selection approaches are based on gene ranking (filter approach). In these gene ranking approaches, each gene is evaluated individually and assigned a score which approximates the relative strength of the gene. Genes are then ranked by their scores and top-ranked genes are selected. Golub [5] used correlation measure to determine relationship between expression levels in samples with its class label to select top genes. In literature [6, 7, 8] many approaches are suggested that adopt the same principle. The problem with ranking methods is that selected genes are often highly correlated [2, 9]. Besides being an additional computational burden, this redundancy can also skew the result and may lead to misclassification. Also, the selected subset so obtained may not perform well on a given classifier as we are not using a classifier while selecting such subset. It is observed that wrapper approaches, which involve a classifier, perform better on small and medium dimensional data. However, they cannot be applied directly on high dimensional microarray dataset as they require huge computation. We can overcome this problem by determining a smaller set of genes for wrapper approach. This is possible if we can group correlated or similar genes into clusters and then select a representative gene from each cluster. The collection of these representative genes of each cluster can provide us a reduced set of independent genes. Wrapper method can then be applied to this reduced set of genes to get a set of discriminatory genes for better classification.

Clustering is the task of assignment of a set of objects into clusters so that objects in the same cluster are similar in some sense. In literature many diverse clustering techniques have been used for grouping such correlated or similar genes. Self-organized maps (SOM) [10], hierarchical [11] and K-means clustering [12] are some of the most widely used clustering techniques. Each technique is associated with certain advantages and disadvantages.

An important step in any clustering technique is to select a distance or similarity measure between two objects. In literature, a large number of similarity or distance measures have been used to determine similarity between two genes. Some of the commonly used similarity measures are Euclidean distance and Pearson correlation coefficient. However, Euclidean distance is not suitable to capture functional similarity such as positive and negative correlation, and interdependency [13]. Also Euclidean distance is suitable only for a data which follows a particular distribution [14]. Pearson coefficient is not robust to outliers and it may assign a high similarity score to a pair of dissimilar genes [15]. Both these measures are also sensitive to scaling and rotation. Maximal information compression index is suggested in literature [16] for measuring redundancy between two features. The maximal information compression index $\lambda_2(x_1, x_2)$ for two random variables x_1 and x_2 is defined as

$$\lambda_2(x_1, x_2) = \frac{\sigma_1 + \sigma_2 - \sqrt{(\sigma_1 + \sigma_2)^2 - 4\sigma_1\sigma_2(1 - \rho(x_1, x_2))^2}}{2} \tag{1}$$

where σ_1, σ_2 are the variance of x_1 and x_2, respectively and $\rho(x_1, x_2)$ is the correlation between x_1 and x_2.

The value of λ_2 measures dependency between x_1 and x_2 and is zero when the features are linearly dependent. It increases as the amount of dependency decreases. The measure λ_2 possesses several desirable properties such as symmetry, sensitivity to scaling and invariance to rotation. Some of these properties are not present in the commonly used Euclidean distance and correlation coefficient. Hence λ_2 may be a good choice for measuring similarity or redundancy between the two features. We have used different clustering techniques in our experiments. A brief description of each is given below.

2.1 Divisive Hierarchical Clustering

It is a class of clustering algorithms based on top-down strategy. In this, one starts with all objects in one cluster [17] and subdivides the cluster into smaller clusters until it satisfies certain termination condition, such as a desired number of clusters are obtained or the diameter of each cluster is within a certain threshold. For hierarchical clustering three parameters are needed. Firstly, some distance measures for evaluating the similarity between different objects. Secondly, a way for choosing the cluster to be splitted, and last but not the least a heuristics for splitting a given cluster into two or more clusters.

In our algorithm we have used maximal compression index for measuring the similarity between two genes. Since we are interested in clusters with maximum redundancy, i.e. each cluster should have maximum redundancy so representative entropy measure is used for selecting a cluster to be splitted from a set of available clusters. Representative entropy [16] measures the amount of redundancy among genes in a given cluster. For a cluster containing p genes with covariance matrix Σ, representative entropy, H_R of a cluster is given by

$$H_R = -\sum_{l=1}^{p} \overline{\lambda_l} \log \overline{\lambda_l} \tag{2}$$

where $\overline{\lambda_l} = \frac{\lambda_l}{\sum_{l=1}^{p} \lambda_l}$ and $\lambda_l, l = 1, 2, \ldots, p$ are the eigenvalues of the matrix Σ.

H_R attains a minimum value (zero) when all the eigenvalues except one are zero, or in other words when all the information is present along a single direction. If all the eigenvalues are equal, i.e. information is equally distributed among all the genes, H_R is maximum. High value of representative entropy represents low redundancy in the cluster. Since we are interested in partitioning the original subspace into homogeneous clusters, each cluster should have low representative entropy. So we split a cluster which has maximum H_R among a given set of clusters as it contains more non-redundant genes.

For splitting a given cluster into two clusters two different strategies are used. The first strategy is based on maximal compression index and the second strategy is based on graph theoretic approach (NCUT). A brief explanation of each of them is given below.

2.1.1 Partition Based on Maximal Compression Index

In this strategy, the maximal compression index between all the genes in the cluster is calculated. Then two genes with maximum dissimilarity, i.e. maximal compression index, are chosen and are made the centres of two newly formed clusters. The remaining genes are then assigned to one of the clusters depending upon its distance from the centres of newly formed clusters. The outline of this clustering technique is given below:

Algorithm for Divisive hierarchical clustering based on Maximal Compression Index

```
Initialization: Set $C$ = initial set of genes;
$S$ = empty set /* Set of Selected Attributes */
Choose Cluster_Size
Calculate dissimilarity Matrix $W$ using maximal information compression index
Choose two genes with maximum dissimilarity and make them the centres of
    two newly formed clusters $C_1$ and $C_2$. // finding centres of new clusters
Each gene in the original cluster is assigned to one of the clusters $C_1$ and $C_2$
    depending upon their similarity with the center. // Splitting original cluster
    into cluster $C_1$ and $C_2$.
no_of_clusters = 2;
While (no_of_clusters $\leq$ Cluster_Size)
Begin
  For each cluster calculate the representative entropy $H_R$
  Choose the Cluster $C_i$ having maximum entropy.
  Choose two genes in $C_i$ with maximum dissimilarity and make them the centres
    of two newly formed clusters $C_{i1}$ and $C_{i2}$. // finding centres of new clusters
  Each gene in the original cluster is assigned to one of the clusters $C_{i1}$ and $C_{i2}$
    depending upon their similarity with the center. // Splitting original cluster
    into cluster $C_{i1}$ and $C_{i2}$.
  no_of_clusters = no_of_clusters + 1
End
```

2.1.2 Partition Based on NCUT

The research work [18, 19] proposed an efficient normalized cut (NCUT) method based on graph theoretic approach for image segmentation. The normalized cut criterion measures both the total dissimilarity between the different groups as well as total similarity within the groups. This can also be used for clustering of correlated genes in microarray data. In NCUT a given graph $G = (V, E)$, where $v_i \in V$ represents a gene and $e(v_i, v_j) \in E$ represents similarity between two genes v_i and v_j is divided into two disjoint sets A and B. For partitioning of the genes into two sets A and B, the capacity of the normalized cut ($Ncut$) is defined as

$$Ncut(A, B) = \frac{cut(A, B)}{assoc(A, V)} + \frac{cut(A, B)}{assoc(B, V)} \tag{3}$$

where

$$cut(A, B) = \sum_{u \in A, v \in B} w(u, v) \text{ and } assoc(A, V) = \sum_{u \in A, t \in V} w(u, t). \tag{4}$$

To determine a better partition of a cluster the value of $Ncut$ should be minimized. Determining minimum normalized cut is a NP-hard problem. Shi and Malik [18] have shown that this problem can be reformulated as eigenvalue problem which is given by

$$D^{-\frac{1}{2}}(D-W)D^{-\frac{1}{2}}x = \lambda x. \tag{5}$$

It has been shown by Shi and Malik [18] that second smallest eigenvector of the above generalized eigenvalue system is the real valued solution to our minimum normalized cut problem. Hence, the second smallest eigenvector can be used to partition the original cluster into two clusters. The outline of NCUT clustering technique is given below:

Algorithm for NCUT Clustering

```
1. Initialization: Set $G$ = initial set of genes;
2. $S$ = empty set /* Set of Selected Attributes */
3. Choose Cluster_Size
4. Calculate the Similarity Matrix $W$ using Maximal information compression index
5. Define $D$ where $D(i) = \sum_j W(i,j)$
6. Solve eigenvalue problem $D^{-1/2}(D-W)D^{-1/2}x = \lambda x$
7. Use the eigenvector with second smallest eigenvalues to divide the original cluster
   $C$ into two clusters.
8. no_of_clusters = 2;
9. While (no_of_clusters $\leq$ Cluster_Size)
10. Begin
11.         For each cluster calculate the representative entropy $H_R$
12.         Choose the Cluster $C_i$ having the maximum entropy
13.         Repeat step (4)–(7) for Cluster $C_i$
14.         no_ of_ clusters = no_ of_ clusters + 1;
15. End
```

2.2 NMF Clustering

Non-negative matrix factorization (NMF) is a group of algorithms in multivariate analysis and linear algebra where a matrix X, consisting of only positive values, is factorized into (usually) two matrices F and G such that $X = FG^T$. It can be traced back to 1970s and is studied extensively by Paatero [20]. The work of Lee and Seung [21, 22] brought much attention to NMF in machine learning and data mining fields. It has proven to be a powerful technique for dimension reduction and clustering. A recent theoretical analysis [23] shows the equivalence between

NMF and k-means clustering. Given a data set in matrix $X(n \times m)$ containing m samples in n-dimensional space, in which each entry is nonnegative, NMF finds an approximation as

$$X \sim FG^T \tag{6}$$

where F is $n \times d$ and G is $m \times d$ matrices. This is solved by finding F and G such that

$$\min_{F,G} |X - FG^T|^2, \text{ s.t.} F \geq 0, G \geq 0. \tag{7}$$

Generally speaking, for any given solution (F, G) of NMF: $X \sim FG^T$ there exits large number of matrices (A, B) such that

$$AB^T = I, FA \geq 0 \text{ and } GB \geq 0. \tag{8}$$

Thus (FA, GB) is also the solution with the same residue $\min |X - FG^T|^2$. So orthogonality condition is imposed on matrix to ensure uniqueness of the solution [24]. Therefore, Equation (7) reduces to

$$\min_{F,G} |X - FG^T|^2, \text{ s.t.} F \geq 0, G \geq 0, G^T G = I. \tag{9}$$

An iterative update algorithm was given by Ding [24] for updating F and G. According to this F and G can be updated using the following equations:

$$G_{jk} \leftarrow G_{jk} \frac{(X^T F)_{jk}}{(GG^T X^T F)_{jk}} \tag{10}$$

$$F_{ik} \leftarrow F_{ik} \frac{(XG)_{ik}}{(FG^T G)_{ik}}. \tag{11}$$

In application of cluster analysis, clustering is done based on the value of G. It assigns sample j to cluster i if G_{ji} is the largest element in row j of G.

3 PROPOSED ALGORITHM

In our proposed algorithm, first the original gene set is divided into some distinct clusters such that the genes within a cluster are highly correlated to each other while those in different clusters are independent. For clustering three clustering techniques described in Section 2 are employed. For measuring the distance between genes maximal compression index is used. After the genes are clustered, the best gene from each cluster is selected using t-statistics to create a set of independent genes. This set of independent genes may contain genes which are not relevant for classification. Therefore to select a set of relevant genes, Sequential Forward Selection (SFS) is applied. In this, one starts with an empty set and then adds one feature at a time that increases a given criterion. Our aim is to select a set of discriminatory genes giving maximum accuracy. Hence, the criterion chosen for SFS is accuracy of the classifier. As the number of samples is much lower in microarray datasets, leave-one-out cross-validation is used for calculating accuracy. In Leave-one-out cross-validation (LOOCV) the training set containing n samples is divided into n disjoint sets each containing a single sample. The classifier is trained n times, each time with a different set (containing single sample) held out as a validation set. The estimated accuracy of Leave-one-out cross validation is the mean of these n accuracies. The outline of the proposed algorithm is given below:

Algorithm for Proposed Gene Selection Method PHASE 1

```
1. Initialization: Set $G =$ initial set of genes;
2. Choose Cluster_Size $n$
3. Divide $G$ into $n$ clusters $C_1, C_2, \ldots, C_n$ using any one of the techniques described
   in Section 2.
4. $S =$ empty set
5. For $i = 1$ to $n$
6.         Find the informative gene $g_i$ from cluster $C_i$ using $t$-statistics.
7.         $S = SUg_i$;
8. end
```

PHASE II // to determine subset of genes which provides maximum classification accuracy

```
1. Initialization $R = \Phi$
2. For each calculate classification accuracy (CA) for classifier $M$.
3. $[x_k, max\_acc] = \max_i \mathrm{CA}(x_i)$
4. $R = R \cup x_k; S = S - x_k$; R_min $= R$ // R_min is the gene subset corresponding
   to maximum accuracy
5. For each $x_j \in S$ calculate classification accuracy of $S \cup x_j$ for classifier $M$
6.         $[x_k, max\_acc] = \max_i \mathrm{CA}(S \cup x_i)$
7.         $R = R \cup x_k; S = S - x_k$;
8.         if new_max_acc > max_acc then R_min = $R$; max_acc = new_max_acc;
9. Repeat steps 5–8 until max_acc = 100 or $S = \Phi$
10. Return R_min, max_acc
```

4 EXPERIMENTAL SETUP AND RESULTS

To test the efficacy of our proposed algorithm we have performed experiments on six publicly available datasets. The details of these datasets are given below.

The Colon Tumor Dataset: This dataset was first used by Alon et al. [25]. It contains 62 samples collected from colon-cancer patients. Out of 62 samples, 40 samples are collected from patients suffering from colon tumor and 22 samples are of healthy patients. In total, there are 6 500 genes but in most of the experiments 2 000 genes with the highest minimal intensity are used [25]. This dataset is downloaded from Kent Ridge Bio Medical repository.

The Ovary Cancer Dataset: This dataset was first used by Emanuel F. Petricoin III [26]. It is a two class problem. It contains 253 samples among which 91 samples are of healthy persons and 162 samples are of persons suffering from ovarian cancer. The entire Ovarian Cancer Dataset contains expression information of 15 154 genes. This dataset is downloaded from Kent Ridge Bio Medical repository.

The Leukemia Dataset: This dataset was first used by Scott A. Armstrong [27]. It is referred as ALL-AML-3. It consists of gene expression of three cases: acute myeloid leukemia (AML), acute lymphoblastic leukemia T-cell (ALL-T) and acute lymphoblastic leukemia B-cell (ALL-B). It contains 72 samples out of which 24 are ALL, 20 are MLL and 28 are AML samples. The entire leukemia data set contains the expression information of 7 129 genes. The data is downloaded from `http://www-genome.wi.mit.edu/cgi-bin/cancer/publications`.

The SRBCT Dataset: This dataset was first used by Khan [28]. SRBCT is the dataset of small round blue cell tumor. This group of highly malignant neoplasms accounts for approximately 10 % of all solid tumors to affect children under the age of 15 years based on incidence. They are generally composed of small round cells that appear blue when stained by conventional histopathological processes. Owing to their morphological similarities, unambiguous clinical diagnosis is difficult. The expression dataset for the SRBCT includes four types of cancers, neuroblastoma (NB), rhabdomyosarcoma (RMS), Burkitt lymphoma (BL), and the Ewing family of tumours (EWS). The SRBCT dataset contains 83 samples out of which 29 are EWS samples, 11 are BL samples, 18 are NB samples and 25 are RMS samples. The entire dataset includes the expression data of 2 308 genes. It is available at `http://research.nhgri.nih.gov/microarray/supplement/`.

The Prostate Dataset: This dataset was first used by Singh [29]. The Prostate dataset contains 102 samples out of which 52 samples are samples of person suffering from prostate tumor and 50 samples are of healthy persons. The original dataset contains 12 600 genes. The dataset is pre-processed. The intensity thresholds are set at 10 − 16 000 units; then the genes with max/min ≤ 5 or $(\max - \min) \leq 50$ are filtered out. After pre-processing, a dataset with 102 samples and 5 966 genes is obtained.

The LungCancer dataset: This dataset was first used by Gavin J. Gordon [30]. It is used to distinguish between malignant pleural mesothelioma (MPM) and adenocarcinoma (ADCA) of the lung. There are 181 tissue samples out of which

31 are of MPM and 150 are of ADCA. Each sample is described by 12 533 genes. This dataset is downloaded from Kent Ridge Bio Medical repository.

Datasets are first normalized using Z score. During the first phase of the algorithm the original gene set is partitioned into k clusters using a clustering algorithm. We have used three different clustering techniques described in Section 3. Once clustering is complete, from each cluster the most relevant gene is selected using t-statistics. Thus a pool of k independent genes is created. The experiment is conducted for different cluster sizes (k). The cluster sizes are taken as 30, 40, 50 and 60. During the second phase Sequential Forward Feature Selection method is applied to obtain a set of relevant genes which provides maximum classification accuracy. Classification accuracy (Leave-one-out cross validation) of the classifier is used as a criterion in forward feature selection. The different classifiers used in our experiments are linear discriminant classifier (LDC), quadratic discriminant classifier (QDC), k-nearest neighbor (KNN) and support vector machine (SVM). For KNN the optimal value of k is chosen. In SVM linear kernel is used.

Table 1 shows the best results obtained among various cluster sizes by different clustering techniques for different classifier. Following observations are made from Table 1:

1. For Colon dataset, with KNN maximum accuracy of 100 % is achieved with 8 genes using NMF. With QDC a maximum accuracy of 98.38 % with 25 genes using NMF. For LDC maximum accuracy of 98.38 % is achieved with 32 genes and for SVM an accuracy of 96.77 % is achieved with 19 genes using NCUT.
2. For SRBCT dataset, maximum accuracy of 100 % is achieved for all the classifiers with genes selected by different clustering algorithms. Best result obtained is 100 % accuracy with 5 genes for SVM using NCUT.
3. For Leukemia dataset, accuracy of 100 % is achieved for all the classifiers. Best result is 100 % accuracy with 5 genes for QDC and KNN using NCUT.
4. For Prostate dataset, accuracy of 100 % is achieved with 5 genes for SVM using NMF. With LDC and QDC maximum accuracy of 98.03 % is achieved using NMF. For KNN we are able to achieve maximum accuracy of 99.01 % with 7 genes using NCUT.
5. For LungCancer dataset, accuracy of 100 % is achieved by the genes selected by all the clustering algorithms for all the classifiers. Best result obtained is 100 % accuracy with 2 genes for KNN using NMF.
6. For Ovary dataset, accuracy of 100 % is achieved by using all clustering algorithms for all the classifiers. Best result obtained is 100 % accuracy with 2 genes for KNN using NCUT.
7. For all the datasets except Colon and Prostate, for each classifier we are able to find a set of genes which gives 100 % LOOCV accuracy.

8. None of the clustering technique seems to be clear winner. Performance of NMF and NCUT methods is better compared to hierarchical clustering. It has also been observed that performance of NMF and NCUT is comparable.
9. Gene selection depends on the choice of the classifier used.

Dataset	Classifier	Clustering Method		
		Hierachial	NCUT	NNMF
Colon	LDC	93.54 (6)	98.38 (32)	91.93 (11)
	QDC	95.16 (6)	95.16 (6)	98.38 (25)
	KNN	98.38 (12)	96.77 (11)	100 (8)
	SVM	95.16 (19)	96.77 (19)	95.16 (31)
SRBCT	LDC	100 (7)	100 (31)	100 (14)
	QDC	100 (18)	97.59 (11)	100 (11)
	KNN	100 (6)	100 (6)	100 (7)
	SVM	100 (6)	100 (5)	100 (8)
Leukemia	LDC	97.22 (5)	100 (7)	100 (6)
	QDC	100 (8)	100 (5)	100 (8)
	KNN	100 (22)	100 (5)	100 (6)
	SVM	100 (7)	100 (46)	100 (32)
Prostate	LDC	97.05 (36)	97.05 (5)	98.03 (5)
	QDC	98.03 (9)	97.05 (19)	98.03 (8)
	KNN	99.01 (3)	99.01 (7)	98.03 (5)
	SVM	99.01 (19)	99.01 (15)	100 (5)
LungCancer	LDC	100 (4)	100 (3)	100 (4)
	QDC	100 (3)	100 (4)	100 (3)
	KNN	100 (2)	100 (3)	100 (2)
	SVM	100 (3)	100 (3)	100 (4)
Ovary	LDC	100(6)	100 (3)	100 (3)
	QDC	100 (3)	100 (3)	100 (3)
	KNN	100 (4)	100 (2)	100 (3)
	SVM	100 (4)	100 (3)	100 (3)

Table 1. Comparison of best results obtained by different clustering techniques

In Table 2, we have compared the performance of our proposed method in terms of classification accuracy and the number of genes with some already existing gene selection methods in literature [6, 28, 31, 32, 33, 34, 35, 36, 37, 38]. It can be observed from Table 2 that the performance of our proposed algorithm is significantly better in terms of both classification accuracy and number of genes selected.

The smallest gene subsets giving the maximum accuracy for different datasets are listed in Table 3. We have also compared the time taken by different clustering algorithms. Table 4 shows the time taken (in seconds) by different clustering algorithms when cluster size is 30. It can be observed from Table 4 that the time taken by NMF clustering is significantly less compared to NCUT and hierarchical clustering.

COLON		OVARY	
Proposed method	100 (8)	Proposed Method	100 (2)
PSO + ANN [6]	88.7	PSO + ANN [6]	97.0
Yuechui and Yaou [38]	90.3	NB [33]	96.2
BIRSW [35]	85.48 (3.50)	BKS [33]	97.0
BIRSF [35]	85.48 (7.40)	DT [33]	97.8
PROSTATE		LUNGCANCER	
Proposed Method	100 (5)	Proposed Method	100 (2)
GAKNN [34]	96.3 (79)	GS2 + KNN [6]	93.1 (44)
BIRS [35]	91.2 (3)	GS1 + SVM [6]	98.6 (4)
		Chos + SVM [6]	98.6 (80)
		Ftest + SVM [6]	98.6 (94)
		Shah and Kaushik [36]	100 (8)
		PSO + ANN [6]	98.3
		Yuechui and Yaou [38]	98.3
LEUKEMIA		SRBCT	
Proposed Method	100 (5)	Proposed Method	100 (5)
GS2 + KNN [6]	98.6 (10)	GS2 + SVM [6]	100 (96)
GS1 + SVM [6]	98.6 (4)	GS1 + SVM [6]	98.8 (34)
Chos + SVM [6]	98.6 (80)	Chos + SVM[6]	98.8 (80)
Ftest + SVM [6]	98.6 (33)	Ftest + SVM [6]	100 (78)
Fu and Liu [31]	97.0 (4)	Fu and Liu [31]	100 (19)
Guyon [32]	100 (8)	Tibsrani [37]	100 (43)
Tibsrani [37]	100 (21)	Khan [28]	100 (96)

Table 2. Comparison of Maximum Classification accuracy and number of genes selected with other state of art methods

To check the generalization capabilities of the genes selected by our proposed algorithm, we also calculated 10-fold cross validation using the different gene subsets selected by our algorithm. The results of the same are given in Table 1. Tables 1 and 5 show that classification accuracy obtained by 10-fold cross validation is approximately the same as the classification accuracy obtained by LOOCV for all the datasets and classifiers except colon dataset with QDC. This shows that the selected gene subsets are good for building classifier.

Dataset	Classifier	LOOCV	Gene Set (Gene Accession Number)
Colon	KNN	100 %	(M26383, X86693, D14812, R44418, T41204, M64231, H55916, M81651)
SRBCT	SVM	100 %	(796258, 812105, 770394, 207274, 183337)
Leukemia	SVM	100 %	(M23197_at, U05259_rnal_at, J04132_at, HG2036_at HT2090_at, U85611_at)
Prostate	SVM	100 %	(37639_at, 2041_i_at, 41504_s_at, 32137_ at, 38259_ at)
LungCancer	KNN	100 %	(33328_at, 1818_s_at)
Ovary	KNN	100 %	(MZ2.7921478, MZ245.24466)

Table 3. Gene Subset giving maximum accuracy for different datasets

Dataset	Time in Seconds		
	Hierarchical Clustering	NCUT	NMF
Colon	1 758.3	1 470.8	23.6
SRBCT	2 320.4	1 940.4	69.8
Prostate	3 150.5	2 529.4	268.7
Leukemia	4 024.7	3 395.6	467.8
LungCancer	8 996.8	8 256.7	748.2
Ovary	14 678.9	12 568.8	1 289.6

Table 4. Time taken by different clustering techniques when cluster size is 30

5 CONCLUSION

In this paper, we have proposed a two stage algorithm for finding a set of discriminatory genes for classifying microarray datasets. Microarray datasets contain a large number of redundant and irrelevant genes. Since microarray datasets contain many correlated genes, the proposed approach first reduces redundancy by grouping correlated genes. We have used a different similarity measure, namely maximal information compression index which has not been used before for microarray datasets. A representative gene from each cluster is selected using t-statistics. The size of this set obtained is small. This allows us to use wrapper approach at the second stage. The use of wrapper method at the second stage gives a better subset of genes. Expe-

Dataset	LDC	QDC	KNN	SVM
Colon	91.94 ± 2.28	77.10 ± 3.85	99.03 ± 0.88	97 ± 0.88
SRBCT	99.28 ± 1.08	98.55 ± 1.32	99.04 ± 1.01	100 ± 0
Leukemia	99.44 ± 0.76	100 ± 0	99.72 ± 0.62	98.89 ± 0.62
Prostate	97.25 ± 0.44	97.65 ± 0.69	97.65 ± 1.32	99.22 ± 0.44
LungCancer	99.56 ± 0.25	100 ± 0	99.45 ± 0.39	100 ± 0
Ovary	100 ± 0	100 ± 0	99.84 ± 0.22	99.60 ± 0

Table 5. Results of 10 fold Cross validation using the best gene subset selected

rimental results show that our proposed method is able to achieve a better accuracy with a small number of genes. In the first stage we have investigated three different techniques, namely divisive hierarchical clustering based on maximal compression index, divisive clustering technique based on NCUT and NMF for clustering. In the second stage we have used four different classifiers – SVM, KNN, LDC and QDC. None of the clustering algorithms or classifiers seems to be a clear winner but for each dataset we are able to find a set of genes that gives 100 % LOOCV accuracy for some classifier. Comparisons with other state-of-the-art methods show that our proposed algorithm is able to achieve better or comparable accuracy with lower number of genes for all the datasets used.

REFERENCES

[1] Bellman, R: Adaptive Control Processes. A Guided Tour, Princeton University Press, 1961.

[2] Guyon, I.—Elisseeff, A.: An Introduction to Variable and feature Selection. Journal of Machine Learning Research, Vol. 3, 2003, pp. 1157–1182.

[3] Wang, Y.—Fillia, S. M.—James, C. F.—Justin, P.: HykGene: A Hybrid Approach for Selecting Marker Genes for Phenotype Classification Using Microarray Gene Expression Data. Bioinformatics, Vol. 21, 2005, No. 8, pp. 1530–1537.

[4] Iffat, A.—Leslie, G.—Smith, S.: Feature Subset Selection in Large Dimensionality Domains. Pattern Recognition, Vol. 43, 2010, No. 1, pp. 5–13.

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

[6] Yang, K.—Cai, Z.—Li, J.—Lin, G. H.: A Stable Gene Selection in Microarray Data Analysis. BMC Bioinformatics, Vol. 7, 2006.

[7] Badli, P.—Long, A. D: A Bayesian Framework for the Analysis of Microarray Expression Data Regularized T-Test and Statistical Inferences of Gene Changes. Bioinformatics, Vol. 17, 2001, pp. 509–519.

[8] Cho, J.—Lee, D.—Park, J. H.—Lee, I. B.: New Gene Selection for Classification of Cancer Subtype Considering Within-Class Variation. FEBS Letters, Vol. 551, 2003, pp. 3–7.

[9] Hanczar, B. et al.: Improving Classification of Microarray Data Using Prototype-Based Feature Selection. SIGKDD Explor., Vol. 5, 2003, pp. 23–30.

[10] Kohonen, T.: Self-Organizing Maps. Springer, Berlin, 1995.

[11] Eisen, M. B.—Spellman, T. P.—Brown, P. O.—Botstein, D.: Cluster Analysis and Display of Genome-Wide Expression Patterns. Proc. Natl. Acad. Sci. USA, Vol. 95, 1998, No. 25, pp. 14863–14868.

[12] Tavazoie, S.—Huges, D.—Campbell, M. J.—Cho, R. J.—Church, G. M.: Systematic Determination of Genetic Network Architecture. Nature Genet, 1999, pp. 281–285.

[13] Jiang, D.—Tang, C.—Zhang, A.: Cluster Analysis for Gene Expression Data: A Survey. IEEE Trans. Knowledge and Data Eng., Vol. 16, 2004, pp. 1370–1386.

[14] Yu, J.—Amores, J.—Sebe, N.—Tian, Q.: Toward Robust Distance Metric Analysis for Similarity Estimation. Proc. IEEE Intl Conf. Computer Vision and Pattern Recognition, 2006.

[15] Heyer, L. J.—Kruglyak, S.—Yooseph S.: Exploring Expression Data: Identification and Analysis of Coexpressed Genes. Genome research, Vol. 9, 1999, pp. 1106–1115.

[16] Mitra, P.—Murthy, C. A.—Pal, S. K.: Unsupervised Feature Selection Using Feature Similarity. IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 24, 2002, No. 3, pp. 301–312.

[17] Han, J.—Kamber, M.: Data Mining: Concepts and Techniques, 2000.

[18] Shi, J.—Malik, J.: Normalized Cuts and Image Segmentation. IEEE Transactions on Pattern analysis and machine Intelligence, Vol. 22, 2000, No. 8, pp. 888–903.

[19] Bala, R.—Agrawal, R. K.—Sardana, M.: Relevant Gene Selection Using Normalized Cut Clustering with Maximal Compression Similarity Measure. Proceedings of the 14$^{\text{th}}$ pacific-Asia Conference on knowledge Discovery and Data Mining, LNCS, Vol. 6119, 2010, pp. 81–88.

[20] Paatero, P.—Tapper, U.: Positive Matrix Factorization: A Non-Negative Factor Model With Optimal Utilization of Error Estimates of Data Values. Environmetrics, Vol. 5, 1994, pp. 111–126.

[21] Lee, D.—Seung, H. S.: Learning the Parts of Objects by Non-Negative Matrix Factorization. Nature, Vol. 401,1999, pp. 788–791.

[22] Lee, D.—Seung, H. S.: Algorithms for Non-Negative Matrix Factorization. Advances in Neural Information Processing Systems, Vol. 13, 2001.

[23] Ding, C.—He, X.—Simon, H.: On Equivalence of Non-Negative Matrix Factorization and Spectral Clustering. Proc. SIAM Data Mining Conf., 2005.

[24] Ding, C.—Li, T.—Peng, W.—Park, H.: Orthogonal Non-Negative Matrix Tri-Factorizations for Clustering. KDD06, August 20–23, Philadelphia, Pennsylvania, USA, 2006.

[25] Alon, U.—Barkai, N.—Notterman, D. A.—Gish, K.—Ybarra, S.—Mack, D.—Levine, A. J.: Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Arrays. PNAS, Vol. 96, 1999, No. 12, pp. 6745–6750.

[26] Petricoin, E. F.—Ardekani, A. M.—Hitt, B. A.—Levine, P. J.—Fusaro, V. A.—Steinberg, S. M.—Mills, G. B.—Simone, C.—Fisherman, D. A.—Kohn, E. C.—Liotta, L. A.: Use of Proteomic Patterns in Serum to Identify Ovarian Cancer. The Lancet, Vol. 359, 2002, No. 9306, pp. 572–577.

[27] Armstrong, S. A.—Staunton, J. E.—Silverman, L. B.—Pieters, R.—Boer, M. L.—Minden, M. D.—Sallan, S. E.—Lander, E. S.—Golub, T. R.—Korsmeyer, S. J.: MLL Translocations Specify a Distinct Gene

Expression Profile That Distinguishes a Unique Leukemia. Nature Genetics, Vol. 30, 2001, pp. 41–47.

[28] Khan, J.—Wei, S.—Ringner, M.—Saal, L. H.—Ladanyi, M.—Westermann, F.: Classification and Diagnosis Prediction of Cancers Using Gene Expression Profiling and Artificial Neural Networks. Nat. Med., Vol. 7, 2001, pp. 673–679.

[29] Singh, D.—Febbo, P. G.—Ross, K.—Jackson, D. G.—Manola, J.—Ladd, C.—Tamayo, P.—Renshaw, A. A.—D'Amico, A. V.—Richie, J. P.—Lander, E. S.—Loda, M.—Kantoff, P. W.—Golub, T. R.—Sellers, W. R.: Gene Expression Correlates of Clinical Prostate Cancer Behavior. Cancer Cell, Vol. 1, 2002, pp. 203–209.

[30] Gordon, G. J.—Jensen, R. V.—Hsiao, L. L.—Gullans, S. R.—Blumenstock, J. E.—Ramasamy, S.—Richards, W. G.—Sugarbaker, D. J.—Bueno, R.: Translation of Microarray Data into Clinically Relevant Cancer Diagnostic Tests Using Gene Expression Ratios in Lung Cancer And Mesothelioma. Cancer Research, Vol. 62, 2002, pp. 4963–4967.

[31] Fu, L. M.—Liu, C. S.: Evaluation of Gene Importance in Microarray Data Based Upon Probability of Selection. BMC Bioinformatics, Vol. 6, 2005, No. 67.

[32] Guyon, I.—Weston, J.—Barnhill, S.—Vapnik, V.: Gene Selection for Cancer Classification Using Support Vector Machine. Machine Learning, Vol. 46, 2003, pp. 263–268.

[33] Hong, J. H.—Cho, S. B.: The Classification of Cancer Based on DNA Microarray Data That Uses Diverse Ensemble Genetic Programming. Artif. Intell. Med., Vol. 36, 2006, pp. 43–58.

[34] Li, L.—Weinberg, C. R.—Darden, T. A.—Pedersen, L. G.: Gene Selection for Sample Classification Based on Gene Expression Data: Study of Sensitivity to Choice of Parameters of the GA/KNN Method. Bioinformatics, Vol. 17, 2001, No. 12, pp. 1131–1142.

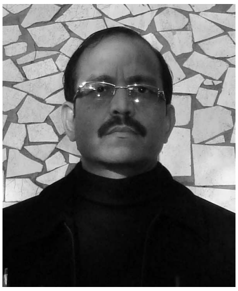
[35] Ruiz, R.—Riqueline, J. C.—Aguilar-Ruiz, J. S.: Incremental Wrapper Based Gene Selection from Microarray Data for Cancer Classification. Pattern Recognition, Vol. 39, 2006, No. 12, pp. 2383–2392.

[36] Shah, S.—Kusiak, A.: Cancer Gene Search with Data Mining and Genetic Algorithms. Computer in Biology Medicine, Vol. 37, 2007, No. 2, pp. 251–261.

[37] Tibsrani, R.—Hastie, T.—Narasimhan, B.—Chu, G.: Diagnosis of Multiple Cancer Types by Shrunken Centriods of Gene Expression. Proc. Natl. Acad. Sci. USA, Vol. 99, 2002, pp. 6567–6572.

[38] Yuechui, C.—Yaou, Z.: A Novel Ensemble of Classifiers for Microarray Data Classification. Applied Soft. Computing, Vol. 8, 2008, pp. 1664–1669.

Rajni Bala is an Associate Professor at Deen Dayal Upadhyaya College, University of Delhi. She has received her Ph. D. in Computers from School of Computer and Systems Sciences, Jawaharlal Nehru University, New Delhi. Her current area of research is pattern recognition.



Ramesh Kumar Agrawal received M. Tech. degree in computer application from Indian Institute of Technology, Delhi. He has done his Ph. D. in computational physics from Delhi University. Presently, he is working as an Associate professor in School of Computer and Systems Sciences, Jawaharlal Nehru University, New Delhi. His current areas of research are classification, feature extraction and selection for pattern recognition problems in domains of image processing, security and bioinformatics.