Original Research Article Allied sciences

International Journal of Pharma and Bio Sciences

ISSN
0975-6299

A NOVEL SUPERVISED GENE CLUSTERING APPROACH BY MINING INTERDEPENDENT GENE PATTERNS

PRADEEP KUMAR MALLICK*[1], DEBAHUTI MISHRA[2], SRIKANTA PATANAIK[3] , AND KAILASH SHAW[4]

[1,2,3,4]Department of Computer Science and Engineering, Siksha 'O' Anusandhan University, Bhubaneswar, Odisha

ABSTRACT

This paper proposes a general methodology of gene clustering based on gene selection to improve the classification accuracy as well as to address the curse of high dimensional datasets. The proposed method first tries to empirically establish the gene clustering technique based on gene selection approach, where the dependency of the genes with respect to clusters have been measured and their categories are defined such as; dependent, independent, lighter dependent and partial dependent within a range of [1-0]. Those genes which fall under the category of lighter dependent [0-0.5] are again checked and get reassigned to the cluster by measuring the interdependency with respect to that cluster and finally all the genes are ranked within clusters. The top most ranked genes of each clusters are taken together to form a pool of genes giving rise to reduced form of dataset. The classification performance of the original datasets and reduced form of those datasets have been measured with mostly used Naïve Bayesian, Decision Tree, Neural Network, Nearest Neighbor and SVM classifiers. Additionally, the classification accuracy of the proposed model has also been verified with few existing feature/attribute selection as well as clustering methods such as; ACA, t-Test, *k*-means, SOM, MRMR etc. An evident finding is that, the proposed algorithm has shown best classification accuracy with excellent predictive capability.

KEYWORDS: Gene selection, Classification, Naïve Bayesian,Decision Tree,Neural Network,Nearest Neighbor,SVM

PRADEEP KUMAR MALLICK
Department of Computer Science and Engineering, Siksha 'O' Anusandhan University, Bhubaneswar, Odisha

***Corresponding Author**

INTRODUCTION

Dealing with huge data generated from microarray data is now a day's a common practical problem. To tackle this many data mining algorithms are being developed. The tools employed are feature selection as preprocessing, clustering, classification etc. To choose from many available methods and the choice of parameters is also to be addressed[1-3] The high dimensionality of microarray datasets with small number of samples and huge number of genes raise a challenge for the microarray data classification. To improve the effectiveness of classifiers, we need to obtain a small representative of datasets by applying feature selection/gene selection technique without losing the property of datasets[4-8] . Existing feature selection/gene selection methods include the removal of non-informative genes which do not contribute to the classification task and the construction of new features which combines the lower level features into higher level features[2,4-6] . The accuracy of the classifiers depends upon the selection of relevant genes. Therefore, in this paper, we tried to propose a gene clustering method based on selection of informative/significant genes by pronouncing the interdependency of the gene patterns among themselves.Our methodology tried to group attributes that are interdependent with each other referred to gene clustering in this study. This gene clustering is able to reduce the original dataset with reduced search space leading to improve the classification accuracy with better predictive ability. This methodology allows discovering the genes with similar expression patterns. The interdependency of the genes has been measured by finding the similar patterns within genes based on mutual information and joint entropy. The degree of belongingness to the clusters is categorized into four categories such as; dependent, independent, lighter dependent and partial dependent within a range of [0-1]. The value 1 represents gene is dependent on the cluster to which it has been assigned, 0 indicates the gene is independent of any cluster, value [>0,0.5] represents gene is lighter dependent on the cluster and the value [>0.5,1] represents the gene is partially dependent on the cluster. Those genes which are categorized as lighter dependent are again reassigned to anyone of the clusters with measuring the degree of membership using fuzzification parameter. The gene interdependency has been computed among the genes with a cluster and once the clusters are formed with a set of disjoint genes, rank of those genes is evaluated with respect to cluster to which they belong to. The top ranked genes of individual clusters are redefined and top most genes are used to form a pool of selected/informative genes.The proposed gene clustering method has been compared with few existing feature/attribute selection as well as clustering algorithms such as Attribute Clustering Algorithm (ACA)[9], t-Test [9-11] , *k*-means[9,12], Self Organizing Map (SOM)[9,13] and MRMR[9,14]. Additionally, the classification accuracy of proposed gene clustering algorithm with selected top ranked genes has been compared with the original datasets. The classifiers used throughout this paper are Naïve Bayesian[9,15-18] Decision Tree[9,19], Neural Network[9, 20-21] , Nearest Neighbor[9,22] and SVM[9, 23-26] . The reminder of the paper is organized as follows: Section 2 discussed the related work addressed by the researches in this area with brief introduction to few existing feature/attribute selection and clustering mechanisms. Section 3 describes the methodologies applied to develop this work; the experimentation and result analysis is given in Section 4 and finally Section 5 concludes the paper.

BACKGROUND STUDY WITH MATERIALS AND METHODS

Traditionally, feature/gene selection research has focused on removing irrelevant and redundant features or attributes. The method of gene selection generally falls into one of the following classes; filter, wrapper, hybrid, embedded. The above mentioned traditional methods can obtain partial good qualities of genes but takes more time. To address this, many heuristic, meta-heuristic optimization methods have been introduced 27-29 This section focuses of recent developments in feature/attribute selection mechanisms.In case of gene expression data, gene grouping and selection are important preprocessing steps for many data mining algorithms. Wai-Ho Au *et al.*[9] has proposed an attribute clustering method (ACA) which is used to group genes based on their interdependence. Significant genes selected from each group then contain useful information for gene expression classification and identification. Hala M. Alshamlan *et al.*[14] proposed a new hybrid gene selection method namely Genetic Bee Colony algorithm which combines GA with Artificial Bee Colony algorithm. In case of text categorization many feature selection methods has been used like chi-square statistics, information gain etc but these methods are not reliable for low- frequency terms. In order to solve this drawback Deqing Wang *et al.*[10] proposed a new approach where t-Test is used to measure the diversity of the distribution of a term frequency between the specific category and the entire corpus. Sina Tabakhi *et al* [8]. proposed an unsupervised gene selection method called MGSACO, which incorporates the ant colony optimization algorithm into the filter approach, by minimizing the redundancy between genes and maximizing the relevance of genes. Behrouz Zamani Dadaneh *et al.*[29] proposed an unsupervised probabilistic feature selection using ant colony optimization. Finally, the algorithm looks for the optimal feature subset in an iterative process. Jerome Paul *et al.*[3] introduced two feature selection methods to deal with heterogeneous data which include continues and categorical variables. The proposed method is used to plug a dedicated kernel that handles both kind of variables into recursive feature elimination procedure using a non-linear SVM or Multiple Kernel Learning. Although Support vector data description has been applied to gene selection, it cannot address the problem with multiclass as it only considers the target class. In order to solve this difficulty Jin Cao *et al.*[23] proposed multiple SVD-RFE where a fast feature selection method based on multiple SVDD and applied in multi-class micro-array data and recursive feature elimination scheme is introduced to iteratively remove irrelevant features. B. Chandra *et al.*[21] proposed Spiking Wavelet Radial Basis Neural Network where a new spiking function in the non-linear integrate and fire model and it's inter spike interval is derived and used in the Wavelet Radial Basis Neural Network for classification of gene expression data. The major tasks with gene expression data is to find co-regulated gene groups whose collective expression is strongly associated with sample categories in this regard, Pradipta Majhi and Chandra Das[30] has proposed a gene clustering algorithm to group genes from microarray data. The algorithm is used to find co-regulated genes with strong association to the sample categories, yielding a supervised gene clustering algorithm. The average expression of the genes from each cluster acts as its representative. Some significant representatives are taken to form the reduced feature set to build the classifiers for cancer classification. The mutual information is used to compute both gene-gene redundancy and gene-class relevance. Milos Krejnik and Jiri Kiema[31] has analyzed the applicability of functional clustering for the identification of groups of functionally related genes, here features of biological samples which originally corresponded to genes are replaced by features that correspond to the centroids of gene cluster and then used for classifier learning. Patrick C. H. Ma and Keith C. C. Chan[32] proposed incremental fuzzy mining (IFM). By transforming quantitative expression values into linguistic terms, such as highly or lowly expressed, IFM can effectively capture heterogeneity in expression data for pattern discovery. Based on these patterns, IFM can make accurate gene function predictions and these predictions can be made in such a way that each gene can be allowed to belong to more than one functional class with different degrees of membership. When class information is unavailable, discovering gene expression patterns becomes difficult, to address this problem, Gene P.K Wu *et al.*[33] proposed a new method of 'fuzzifying' the crisp attribute clusters in which first a gene pool with large number of genes are cluster into smaller groups. This study has been motivated from [9-10, 30-31] where, authors tried to group the genes based on distance function, but we have explored the pattern based interdependency among the genes to obtain informative or significant gene clusters.Various classification algorithms are used throughout this paper to empirically establish the performance of proposed model. We have used Naïve Bayesian, Decision Tree, Neural Network, Nearest Neighbor and SVM classifiers. This section discusses the advantages of those classifiers, which motivated us to use for empirical comparison. The advantage of using Naive Bayes is that, it calculates a probability by dividing the percentage of pair wise occurrences by the percentage of singleton occurrences. If these percentages are very small for a given predictor, they probably will not contribute to the effectiveness of the model. Occurrences below a certain threshold can usually be ignored [15-18, 34-36]. The main advantage is interpretability of decision tree is that the acquired knowledge can be expressed in a readable form, easy to interpret, complexity is the down side and biggest benefit is that the output of a decision tree can be easily interpreted (by humans) as rules[19, 37-38]. As on date neural networks are very popular computer tool used for solving lot of different practical problems. First advantage of neural network is network learning; it first, learns results of solved problem and next solves many another similar problems. It is really very comfortable and efficient way of problem solving[20, 39-40]. By employing Nearest Neighbor classifier, we performed experiments to select the best number of neighbors' *k* and the best feature space transformation[22]. SVMs are a new promising non-linear, non-parametric classification technique, which already showed good results in the medical diagnostics, optical character recognition, electric load forecasting and other fields[9, 23-25, 41]. For the purpose of comparison we have applied ACA[9], t-Test [9-10, 35], *k*-means[9,11], SOM[9,13] and MRMR[9,14] on both the datasets and compared with proposed gene clustering algorithm as well as the classification accuracy of original datasets.

METHODOLOGIES FOR EVALUATION

This section discusses the various methodologies adopted for designing a supervised clustering algorithm based on dependency of genes among the cluster centre head for cancer classification and using two benchmark datasets Colon Cancer[42] and Leukemia[43] and also investigates the complexity of proposed algorithm.

(i) Feature Selection

Let $D = \{S, G\}$ be the set of gene expression dataset composed of S samples and G genes. Every row in *D* is characterized by a set of genes $G = \{g_1, g_2, \dots, g_p\}$. Let, D consists of *n* number of samples $S = \{s_1, s_2, \dots, s_n\}$ and each sample, s_r, $r \in \{1, \dots, n\}$ is represented by a vector of *p* genes, with values $s_r = (a_{r1}, a_{r2}, \dots, a_{rp})$, where, $a_{ri} \in dom(g_i)$, $i = 1, \dots, p$.

Definition1

Gene clustering is a process to discover k distinct clusters, $Cl_1, Cl_2, \dots, Cl_k$, by assigning each gene in $\{g_1, g_2, \dots, g_p\}$ to one of these clusters. It can be stated that *gene clustering* is a process such that; $\forall g_i$, $i = 1, \dots, p$, g_i is assigned to Cl_t, $t \in \{1, \dots, k\}$, where $Cl_t \cap Cl_v = \emptyset$ for all $v \in \{1, \dots, k\} - \{t\}$.

The objective of this study is obtain meaningful clusters, in view of this we first experiment the cluster configuration and try to reveal the information about the genes and gene grouping obtained by using proposed method. Therefore, we tried to implement gene clustering which helps in measuring the high correlation or high interdependence genes within a cluster. Most of the clustering algorithms use distance based dissimilarity measurement methods, but proposed work focuses on mining *interdependent patterns* among genes for obtaining significant cluster.

(ii) Measuring Dependence Genes with Cluster

Given *D* dataset of continuous genes can be discretized into finite interval using Optimal Class-Dependent Discretization (OCDD) method proposed by L. Liu *et al.* [44]. After discretizing the domain of the entire gene in *D* can be represented by $dom\,(g_i) = \{u_{i1}, \dots, u_{im}\}, i = 1, \dots, p$. Now assume for *k* distinct clusters there exists *k* centers;

C=$\{C_1, C_2, \ldots, C_k\}$, $C_i = \{c_{i1}, \ldots, c_{ip}\}$, $c_{ij} \in dom(G)$, $\quad i \in 1, \ldots, k$, $j \in 1, \ldots, p$.

Definition 2

The dependency of genes g_i with cluster Cl_j can be computed using equation (1).

$$I(g_i, Cl_j) = \frac{M(g_i, Cl_j)}{J(g_i, Cl_j)} \qquad (1)$$

Where, *M* is the mutual information between gene g_i and cluster Cl_j as computed by equation (2).

$$M(g_i, Cl_j) = \sum_{x=1}^{m_i} \sum_{y=1}^{m_j} P(g_i = u_{ix} \wedge c_j = u_{jy}) log \frac{P(g_i = u_{ix} \wedge c_j = u_{jy})}{P(g_i = u_{ix}) P(c_j = u_{jy})} \qquad (2)$$

Here, c_j is the *Center of Cluster* Cl_j, and $J\ (g_i, Cl_j)$ is the *Joint Entropy* of g_i and Cl_j and given in equation (3).

$$J(g_i, Cl_j) = -\sum_{x=1}^{m_i} \sum_{y=1}^{m_j} P(g_i = u_{ix} \wedge c_j = u_{jy}) \log\left(P(g_i = u_{ix} \wedge c_j = u_{jy}\right) \qquad (3)$$

In equation (2), if $M(g_i, Cl_j) > M(g_i, Cl_z)$, $z \in \{1, \ldots, p\}, z \neq i \neq j$, the dependency of g_i on cluster Cl_j >Cl_z, *M* appears to increase in value with the number of *dom* value of m_i, and m_j, hence it is normalized using *J*, which yields the interdependency among genes with clusters as *I*. The probability of records in D with $g_i = u_{ik}, i \in \{1, \ldots, p\}, k \in \{1, \ldots, m_i\}$ is computed using equation (4).

$$P(g_i = u_{ix}) = \frac{\left|Count_{g_i = u_{ix}}(D)\right|}{m_i} \qquad (4)$$

The joint probability of gene with respect to cluster is computed using equation (5).

$$P(g_i = u_{ix} \wedge c_j = u_{jy}) = \frac{\left|Count_{g_i = u_{ix} \wedge c_j = u_{jy}}(D)\right|}{m_i . m_j} \qquad (5)$$

Where, $i \in \{1, \ldots, p\}, j \in \{1, \ldots, k\}, x \in \{1, \ldots, m_i\}$ and $y \in \{1, \ldots, m_j\}$. The following are inference from equation (1); where, $I(g_i, Cl_j)$ shows the degree of deviation between the gene g_i with cluster Cl_j.

- $I(g_i, Cl_j) = 1$, it signifies that, g_i and Cl_j are dependent,
- $I(g_i, Cl_j) = 0$, signifies g_i and Cl_j are indipendent,
- $0 < I(g_i, Cl_j) \leq 0.5$, signifies g_i and Cl_j are lighter dependent and,
- $0.5 < I(g_i, Cl_j) < 1$, signifies g_i and Cl_j are partial dependent.

(iii) Fuzzifying the clusters

Now the gene g_i which falls in lighter dependency with cluster Cl_j can be checked for degree of membership with other clusters $Cl_t, t \in 1, \ldots, k, t \neq j$. The degree of membership can be calculated using fuzzification given by equation (6).

$$\mu_r(g_i) = \frac{1}{\sum_{c=1}^{k} \left(\frac{GI(g_i, c_j)}{GI(g_i, c_t)}\right)^{\frac{2}{f-1}}}, where\ , f\ is\ fuzzification\ parameter \qquad (6)$$

The fuzzification parameter *f* is a real number >1 for fuzzifying the measure. For *m*=2 means to normalize the measure linearly to make their sum 1, and *m* close to 1 indicates the gene nearby to the center is given more weight than others. Fuzzy membership μ_r defines correlation of each gene of lighter interdependency with the entire group of clusters *r*. *Gene Interdependency* (GI) measures the dependency among the genes with a cluster using equation (7).

$$GI(g_i, c_r) = \sum_{i=1}^{p} I(g_i, g_j), i \in \{1, .., p\},\ c_r = \{g_j | j = 1, \ldots, p \text{ and } i \neq j\} \qquad (7)$$

Now, if $\mu_r(g_i) > \mu_s(g_i)$, $r, s \in 1, \ldots, k, r \neq s$, indicates g_i must belong to cluster *r* rather than cluster *s*. Once the cluster is formed with set of disjoints genes then rank of individual genes can be evaluated with respect to cluster to which they belong to. Top rank genes of individual clusters can help in redefining the new center of clusters and the rank of gene g_i in the cluster c_j can be formulated using equation (8).

$$R(g_i, c_j) = Sort_{index}(GI(g_i, c_j)) \qquad (8)$$

FUNCTIONAL AND STRUCTURAL DESIGN OF PROPOSED GENE SELECTION MODEL

The proposed gene selection model is shown in Figure1. This model works in two phases. In phase1; both the datasets are classified using five well known classifiers such as; Naïve Bayesian, Decision Tree, Neural Network, Nearest Neighbor and SVM and their classification accuracy has been measured and stored for future comparison. In the second phase; the gene selection approach works to rank genes by measuring the interdependency with respect to their clusters and those top ranked genes are selected for creating a pool of genes giving rise to reduced form of datasets. The classification accuracy of those reduced datasets were measured and compared with previously obtained result and also with few existing feature/attribute selection mechanisms. The result analysis part is evident of our proposed model which outperforms other existing models and also the original datasets without losing the classification accuracy. The algorithm of gene selection is also described in this section.

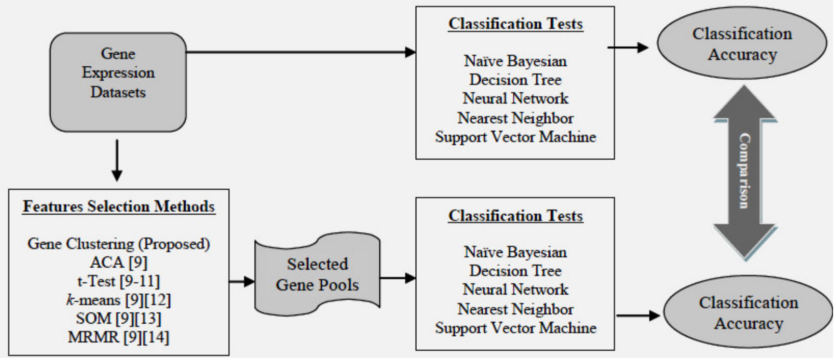


Figure 1
Structural design of proposed gene selection method

Algorithm 1
Gene-Selection

```
Step 1. Initialize number of clusters k, and randomly select k centers C of size p; initialize m.
Step 2. While max_iter or no_change_in_cluster_center repeat step 3 to 7
Step 3. Assign $g_i$ to $Cl_j$ if $I(g_i, Cl_j) > I, i \in \{1, \dots, p\} and\ j \in \{1, \dots, k\}$
Step 4.     For $i = 1\ to\ p$
                For j=1 to k
                    If $I(g_i, Cl_j) > 0$ &&$I(g_i, Cl_j) < 0.5$
                        Compute $\mu_r(g_i)$
Step 5.     if $\mu_r(g_i) > \mu_s(g_i)$ , $r, s \in 1, \dots, k, r \neq s$ than Assign $g_i$ to $Cl_r$ else $g_i$ to $Cl_s$
Step 6.     Evaluate rank of each gene with respect to clusters $R(g_i, c_j)$
Step 7.     Assign top rank gene of each cluster as new center to that cluster.
```

COMPUTATIONAL COMPLEXITY ANALYSIS

Let consider a gene expression dataset composed of *n* samples and *p* gene expression levels. Proposed algorithm requires O (*kp*) operations to assign each gene to a cluster (Step 3). Let us assume that, 50% genes are those whose $0 < I(g_i, Cl_j) < 0.5$ is lighter interdependency, it then performs O($(n/2)^2 pk$) (step 4) comparisons. Let O (*n*) comparisons are made to compute new center for each clusters and *t* be the number of iterations. The computational complexity of proposed algorithm will be O((*kp*+$(n/2)^2 pk$+*n*)*t*) = O$(kn^2 pt)$.

MODEL EVALUATION AND EXPERIMENTAL ANALYSIS

In this section, we empirically evaluate the performance of proposed gene clustering model based on gene selection and classification using MatLab7.10 in Pentium dual core processor on Window10 OS, 2GB of RAM upon two well known benchmark datasets Colon Cancer[42] and Leukemia[43] , those are the same datasets used in 9 and 45-46 with them we have made comparisons with proposed algorithm. The Colon Cancer dataset contains expression levels of 2000 genes and 62 samples from two classes, 40 tumor and 22 normal colon tissue. Leukemia is an affymetrix high-density oligonucleotide array containing 7129 genes and 72 samples from two class leukemia, out of which 47 are acute lymphoblastic leukemia and 25 acute myeloid leukemia.

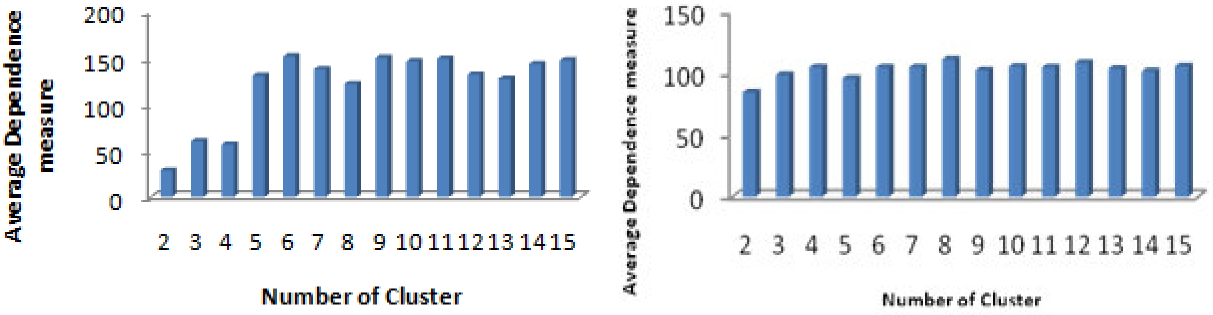


Figure 2
The Average dependence measure over all the clusters found in the (a) colon-cancer and (b) leukemia data sets.

The proposed gene selection methodology first examines the cluster configuration obtained by different methods by measuring the dependencies of genes with respect to clusters, second; study the representative patterns in each cluster from top ranked genes and finally, measuring and comparing the classification accuracy using Naïve Bayesian, Decision Tree, Neural Network, Nearest Neighbor and SVM classifiers of the proposed gene selection model with few existing attribute clustering/feature selection mechanisms such as; ACA, t-Test, *k*-means, SOM and MRMR. In the first set of experiment, the datasets are discretized using OCCD [44] , both the datasets are discretized into two intervals since both the datasets are having two classes. OCCD minimizes the information loss during discretization process and after that we have used interdependency based gene selection method to obtain the clusters of genes[46-47]. The average of the dependency redundancy measure of overall clusters versus the number of clusters formed from the colon-cancer and leukemia is shown in Figure 2. In proposed method, the cluster configuration is formed by measuring the dependency among genes with clusters using equations (1) to (8). From Figure 2, it can be noticed that, the optimal numbers of clusters are six (6) and eight (8) for Colon-Cancer and Leukemia datasets respectively. We can say that, the numbers of clusters obtained are optimal with respect to the average interdependence of genes in clusters. After clustering, top five genes ranked using equation (8) are selected for investigating the representative patters in each cluster as given in Table 1.

Table 1
Selected top five genes in each of the clusters for both datasets (Colon Cancer -6 and Leukemia-8)

Dataset	Clusters	Gene Accession Number				
		Rank1	Rank2	Rank3	Rank4	Rank5
Colon Cancer	1	X51985	X87159	M90684	M16029	T62635
	2	U30498	U19261	R35885	X51346	R77780
	3	R80427	H13292	R73660	X16354	D31887
	5	H72110	L38696	T68848	H40560	T48386
	6	L06328	L16510	J04046	T64941	T40568
Leukemia	1	U51096	M86752	HG1828-HT1857	D50863	M37825
	2	L20859	D31884	U86759_s	L07493	U23028
	3	X13810_s	X63597	L01087	M13485	U90716
	4	X96484	U51240	M93650	M34677	U70671
	5	M29696	M34175	J03242_s	AFFX-HUMRGE/M10098_M	HG2846-HT2983
	6	X83703	U30610	L19297	M28585_f	Z14093
	7	Y09321	U03911	U71087	D13264	HG4460-HT4729
	8	M64595	M81780_cds5	X81882	D49493	X64810

In the second set of experiment, we select the most significant/top ranked genes in each cluster based on average dependency of genes with respect to clusters, fuzzification of genes having lighter dependency based of fuzzification parameters and finally ranking all the genes of each clusters. Moreover, here we tried to discover the coherent nature of the representative genes, their interdependency, similarity, implications of relationship in the formed patterns based on significant top selected genes. Figure 3 and Figure 4 illustrates the most representative gens in Cluster 6 (*k*=6) of Colon Cancer dataset and Cluster 8 (*k*=8) of Leukemia dataset respectively. All the clusters obtained by proposed gene clustering methodology for Colon Cancer is given in Figure 5 and for Leukemia dataset is Figure 6.

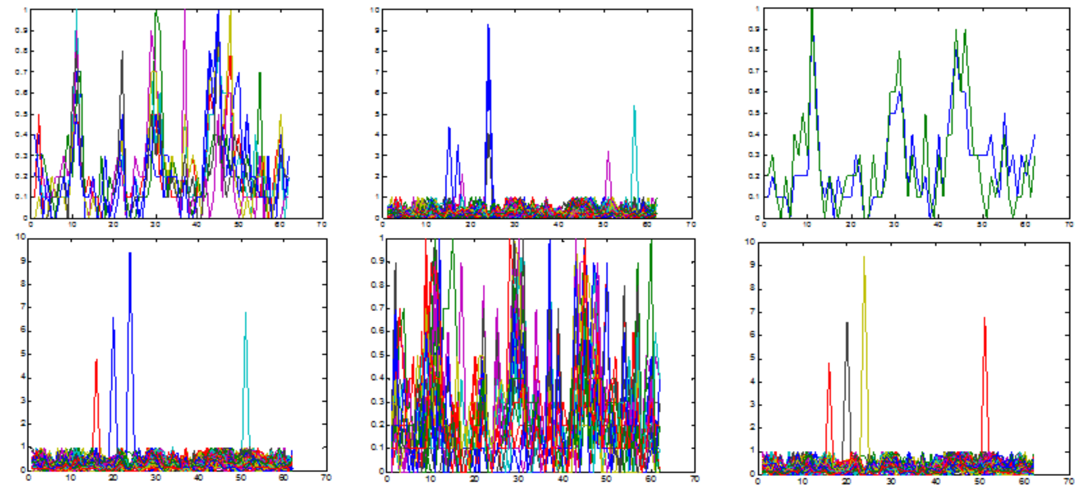


Figure 3
Patterns discovered in cluster k=6 for Colon Cancer dataset

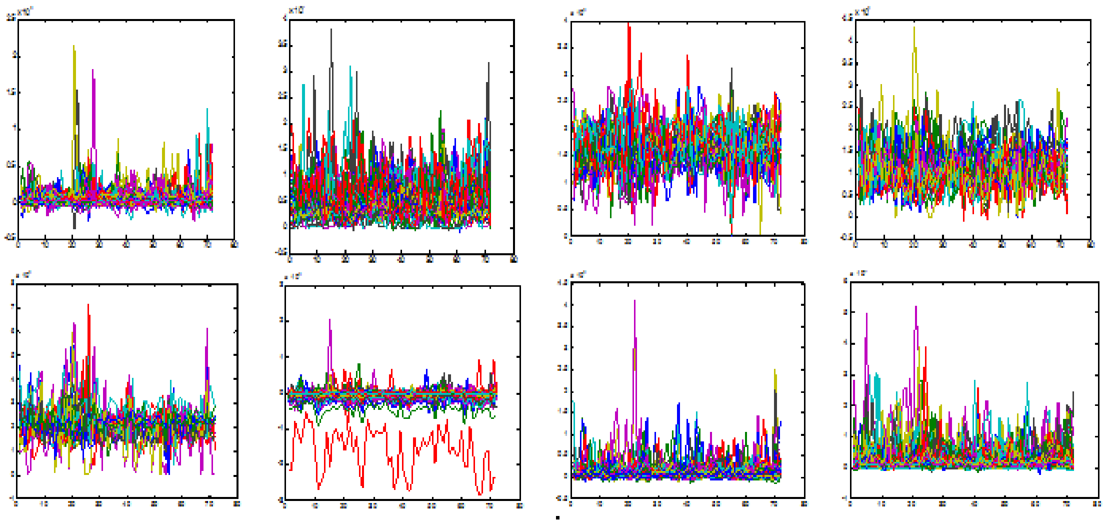
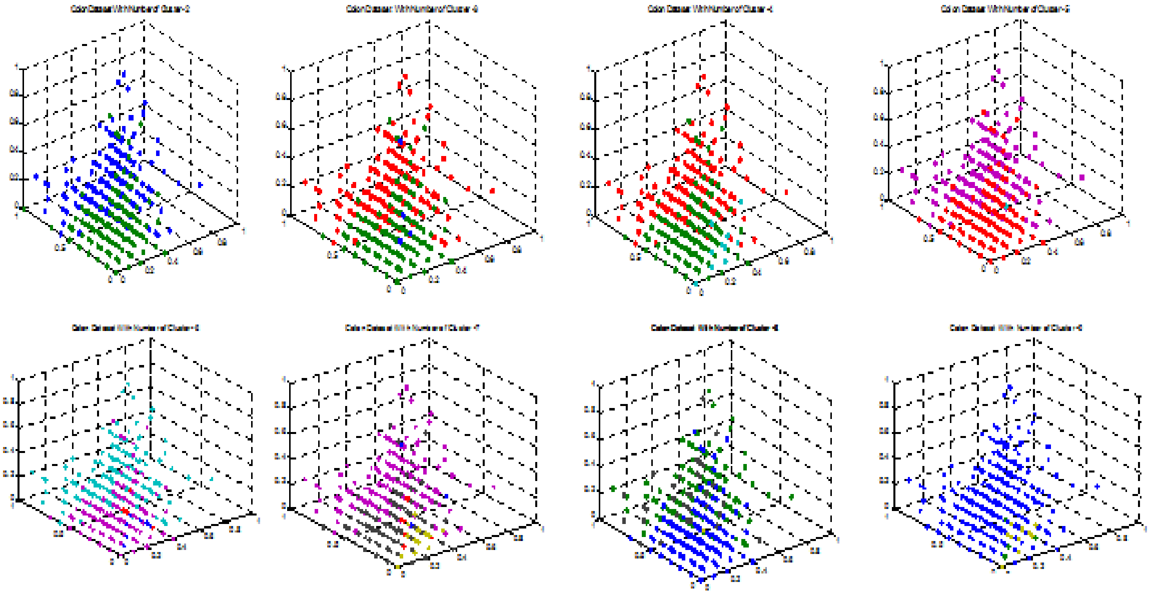


Figure 4
Patterns discovered in cluster k=8 for Leukemia dataset

Clusters obtained by proposed gene clustering method for Colon Cancer dataset



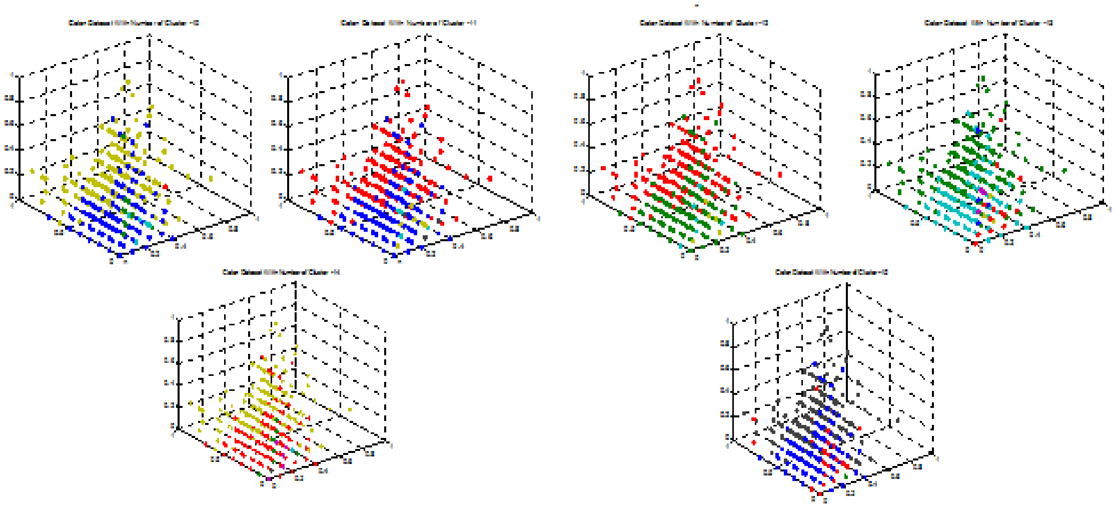


Figure 5
Clusters obtained by proposed gene clustering method for Colon Cancer dataset

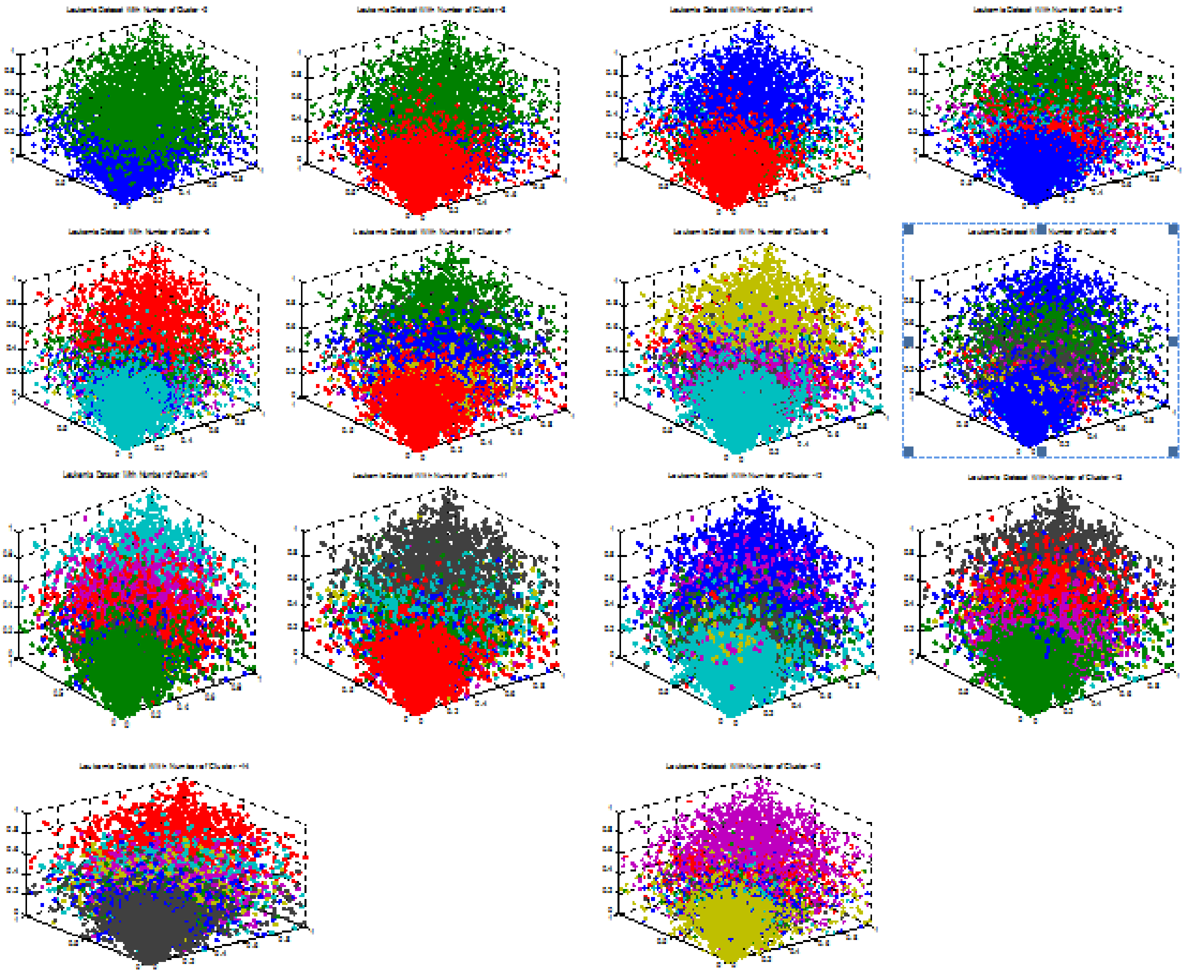


Figure 6
Clusters obtained by proposed gene clustering method for Leukemia dataset

In the third set of experiments, the performance of the proposed gene clustering methodology has been assessed using few classifiers such as; Naïve Bayesian, Decision Tree, Neural Network, Nearest Neighbor and SVM. Class labels of both the datasets were known and this helped us to measure the classification accuracy of the model in comparison with original dataset and reduced dataset containing selected top most genes from each clusters. The classification accuracy measured for both original (not reduced) Colon-Cancer and Leukemia datasets is given in Table 2 and the results obtained by the same set of classifiers on the reduced datasets containing top most ranked genes from each clusters are shown in Table 3, 4, 5, 6 and **7** for Colon Cancer and Table 8, 9, 10, 11 and 12 for Leukemia datasets.

Table 2

The performance of different classification algorithms in colon-cancer and Leukemia datasets (Classification Accuracy in %)

Classifiers	Colon Cancer	Leukemia
Naïve Bayesian	72.58	77.78
Decision Tree	90.2	88.9
Neural Network	83.9	86.2
Nearest Neighbor	79	82.4
SVM	90.87	87.85

The datasets collected are high dimensional datasets, leading to raise the issue of curse of dimensionality. In this paper, we have tried to address this problem by reducing the datasets by selecting only the informative/significant genes. The informative genes can be used to build an accurate classifier by using inductive learning algorithms. Hence, we have selected top *N* genes from each cluster so that a total of 6 × *N* and 8 × *N* genes are selected for *N*=1,…,5, from the Colon Cancer and Leukemia datasets respectively. From the literature [9][15-26] we found that the classifies Naïve Baysian, Decision Tree, Neural Network, Nearest Neighbor, and SVM are mostly used in classifying gene expression datasets. Therefore, in this paper we employed these five classifiers to learn the impact of our selected gene on classification of gene expression dataset. Comparing Table 2 with Table 3, 4, 5, 6 and 7 for Colon Cancer dataset, it is evident that, the classification accuracy obtained after classifying the original dataset is using Naïve Bayesian is 72.58% whereas, the classification accuracy obtained for top 6, 12, 18, 24 and 30 genes is 98.7%, 898.7%, 94.4%, 94.4% 92.2% respectively and also additionally, comparing the reduced form of dataset with top most genes selected with ACA, t-Test, *k*-means, SOM and MRMR feature/attribute selection methods, the proposed method has enhancement of accuracy approximately 15-35%. The classification accuracy observed using Decision Tree classifier is 90.2% for original Colon Cancer dataset, whereas, the reduced form of Colon Cancer has accuracy of 94.3%, 92.6%, 88.2%, 88.2%, 88.2% for top 6, 12, 18, 24 and 30 genes. Additionally, there is an increase of classification accuracy 10-38% while measured with the above mentioned feature/attribute selection mechanisms. The classification accuracy obtained using Neural Network is 83.9% for original Colon Cancer dataset, whereas, the reduced form of Colon Cancer has accuracy of 99.7% for top 6, 12, 18, 24 and 30 genes. Additionally, there is an increase of classification accuracy 3-30% while measured with the above mentioned feature/attribute selection mechanisms. The classification accuracy obtained using Nearest Neighbor is 79% for original Colon Cancer dataset, whereas, the reduced form of Colon Cancer has accuracy of 92.3% for top 6, 12, 18, 24 and 30 genes. Additionally, there is an increase of classification accuracy 2-42% while measured with the above mentioned feature/attribute selection mechanisms. The classification accuracy obtained using SVM is 90.87 % for original Colon Cancer dataset, whereas, the reduced form of Colon Cancer has accuracy of 97.8%, 97.8%, 98.5%, 98.5% and 98.5% for top 6, 12, 18, 24 and 30 genes. Additionally, there is an increase of classification accuracy 7-46% while measured with the above mentioned feature/attribute selection mechanisms.

Table 3
Classification accuracy in % using Naïve Bayesian classifier on top N genes in Colon Cancer dataset.

No. of Genes Selected	Proposed Method	ACA [9]	t-Test[9]	*k*-means [9]	SOM [9]	MRMR [9]
6	**98.7**	88.2	68.3	62.4	58.4	70.3
12	**98.7**	87.8	64.6	54.3	58.4	65.4
18	**94.4**	75.4	64.6	54.3	58.4	65.4
24	**94.4**	75.4	64.6	54.3	60.1	76.7
30	**92.2**	78.3	60.2	54.3	60.1	76.7

Table 4
Classification accuracy in % using Decision Tree classifier on top N genes in Colon Cancer dataset

No. of Genes Selected	Proposed Method	ACA	t-Test	*k*-means	SOM	MRMR
6	**92.4**	82.1	54.1	63.8	50.8	67.5
12	**90.7**	77.5	60.2	63.8	50.8	67.5
18	**87.2**	77.5	55.1	60.2	50.8	67.5
24	**87.2**	77.5	57.4	60.2	50.8	56.5
30	**87.4**	77.5	55.3	60.2	50.8	50.2

Table 5
Classification accuracy in % using Neural Network classifier on top N genes in Colon Cancer dataset

No. of Genes Selected	Proposed Method	ACA	t-Test	*k*-means	SOM	MRMR
6	**99.97**	97.6	70.8	70.1	60.9	88.8
12	**99.97**	94.3	70.8	68.1	60.9	88.6
18	**99.97**	94.3	76.3	62.7	62.2	84.2
24	**99.97**	90.2	76.3	62.7	62.2	84.2
30	**99.97**	90.2	77.1	55.3	62.2	84.2

Table 6
Classification accuracy in % using Nearest Neighbor classifier on top N genes in Colon Cancer dataset

No. of Genes Selected	Proposed Method	ACA	t-Test	*k*-means	SOM	MRMR
6	**92.3**	90.4	82.8	52.5	52.7	66.6
12	**92.3**	92.3	82.8	44.6	51.2	66.6
18	**92.3**	92.3	80.1	44.6	55.4	68.2
24	**92.3**	90.4	80.1	44.6	54.1	68.2
30	**92.3**	90.4	80.1	44.6	54.1	68.2

Table 7
Classification accuracy in % using SVM classifier on top N genes in Colon Cancer dataset

No. of Genes Selected	Proposed Method	ACA	t-Test	*k*-means	SOM	MRMR
6	**97.8**	90.6	77.9	52.4	66.5	66.4
12	**97.8**	90.6	77.9	52.4	67.2	64.4
18	**98.5**	85.2	77.9	51.3	67.2	61.8
24	**98.5**	88.1	75.3	50.7	65.1	61.8
30	**98.5**	88.1	75.3	49.2	65.1	62.2

Similarly, for Leukemia dataset, comparing Table 2 with Table 8, 9, 10, 11 and 12**,** it is observed that, the classification accuracy obtained after classifying the original dataset is using Naïve Bayesian is 77.78% whereas, the classification accuracy obtained for top 8, 16, 24, 32 and 40 genes is 92.7%, 88.6%, 84.2%, 84.2% and 82.4% respectively and also additionally, comparing the reduced form of dataset with top most genes selected with ACA, t-Test, *k*-means, SOM and MRMR feature/attribute selection methods, the proposed method has enhancement of accuracy approximately 8-38%. The classification accuracy observed using Decision Tree classifier is 88.9% for original Leukemia dataset, whereas, the reduced form of Leukemia has accuracy of 92.4%, 90.7%, 87.2%, 87.2% and 87.4% for top 8, 16, 24, 32 and 40 genes. Additionally, there is an increase of classification accuracy 12-40% while measured with the above mentioned feature/attribute selection mechanisms. The classification accuracy obtained using Neural Network is 86.2% for original Colon Cancer dataset, whereas, the reduced form of Leukemia has accuracy of 99.7%, 99.7%, 98.62%, 98.62% and 98.62% for top 8, 16, 24, 32 and 40 genes. Additionally, there is an increase of classification accuracy 1-36% while measured with the above mentioned feature/attribute selection mechanisms. The classification accuracy obtained using Nearest Neighbor is 82.4% for original Leukemia dataset, whereas, the reduced form of Leukemia has accuracy of 95.6% for top 8, 16, 24, 32 and 40 genes. Additionally, there is an increase of classification accuracy 2-45% while measured with the above mentioned feature/attribute selection mechanisms. The classification accuracy obtained using SVM is 87.85 % for original Leukemia dataset, whereas, the reduced form of Leukemia has accuracy of 96.8%, 95.6%, 95.6%, 95.6% and 95.6% for top 8, 16, 24, 32 and 40 genes. Additionally, there is an increase of classification accuracy 2.4-37% while measured with the above mentioned feature/attribute selection mechanisms.

Table 8
Classification accuracy in % using Naïve Bayesian classifier on top N genes in Leukemia dataset

No. of Genes Selected	Proposed Method	ACA	t-Test	*k*-means	SOM	MRMR
8	**92.7**	84.4	58.6	60.1	60.1	69.6
16	**88.6**	65.8	58.6	57.5	60.1	57.9
24	**84.2**	63.4	58.6	57.5	60.1	56.6
32	**84.2**	63.4	58.6	57.5	60.1	57.4
40	**82.4**	63.4	58.6	57.5	60.1	57.4

Table 9
Classification accuracy in % using Decision Tree classifier on top N genes in Leukemia dataset

No. of Genes Selected	Proposed Method	ACA	t-Test	*k*-means	SOM	MRMR
8	**92.4**	80.4	64.4	60.8	61.8	64.6
16	**90.7**	78.8	62.4	60.8	61.8	62.9
24	**87.2**	78.8	62.4	60.8	61.8	60.0
32	**87.2**	76.5	62.4	60.9	57.3	55.1
40	**87.4**	76.5	62.4	60.9	55.2	52.3

Table 10
Classification accuracy in % using Neural Network classifier on top N genes in Leukemia dataset

No. of Genes Selected	Proposed Method	ACA	t-Test	*k*-means	SOM	MRMR
8	**99.97**	98.2	82.4	72.6	65.4	93.4
16	**99.97**	96.7	82.4	72.6	65.4	92.2
24	**98.62**	96.7	84.5	72.6	65.4	94.5
32	**98.62**	92.2	84.5	72.6	65.4	94.5
40	**98.62**	90.4	84.5	72.6	65.4	94.5

Table 11
No. of Genes Selected Nearest Neighbor classifier on top N genes in Leukemia dataset

No. of Genes Selected	Proposed Method	ACA	t-Test	*k*-means	SOM	MRMR
8	**95.6**	93.4	86.2	62.8	55.4	64.3
16	**95.6**	93.4	88.5	55.6	60.3	64.3
24	**95.6**	93.4	88.5	55.6	60.3	65.4
32	**95.6**	93.4	88.5	50.2	58.7	65.4
40	**95.6**	93.4	88.5	49.4	55.4	65.4

Table 12
No. of Genes Selected SVM classifier on top N genes in Leukemia dataset

No. of Genes Selected	Proposed Method	ACA	t-Test	*k*-means	SOM	MRMR
8	**96.8**	93.4	87.9	66.7	64.4	68.3
16	**95.6**	91.4	85.8	57.1	63.3	68.3
24	**95.6**	91.4	82.5	57.7	62.3	65.4
32	**95.6**	90.1	82.5	57.5	62.3	64.4
40	**95.6**	88.4	82.5	49.4	62.3	64.4

As mentioned in the introduction section, the aim of the proposed work is to the find informative set of genes to design a gene clustering model which will lead to improve the classification accuracy efficiently and also to overcome the problem of curse of high dimensionality, the proposed approach addresses all of those. Moreover, the proposed method is essentially different from traditional gene/feature selection approaches. The main contributions are summarized as:

1. A new formulation of gene clustering based on gene selection has been proposed.
2. The dependency of the genes with respect to clusters is computed.
3. The dependency has been categorized to find the significant genes.
4. The optimal fuzzification parameter has been formulated to obtain the gene inter-dependency.
5. Finally, the genes are ranked as per their significance within their clusters and few top most genes are selected from each cluster to form a pool of genes giving ride to reduced form of the original dataset.
6. The classification accuracy, predictive capability and computational complexity are explored by testing with mostly used feature/attribute selection and clustering techniques.

CONCLUSION

This paper presents a gene clustering algorithm based on feature/gene selection for cancer classification. The performance of the model has been evaluated by the predictive accuracy of the Naïve Bayesian, Decision Tree, Neural Network, Nearest Neighbor and SVM classifiers. For both the datasets, significantly better results are found by the proposed gene clustering/gene selection method compared to others. The model is capable of identifying the significant or informative genes that may contribute to reveal the underlying class structure, providing a useful tool for exploratory analysis of biological data.

CONFLICT OF INTEREST

Conflict of interest declared none.

REFERENCES

1. Dessi N, Pes B. Similarity of feature selection methods: An empirical study across data intensive classification task. Expert Systems with Applications. 2015. 42: 4632-4642.
2. Du D, Li K, Li X, Fei M. A novel forward gene selection algorithm for microarray data. Neurocomputing. 2014. 133: 446-458.
3. Paul J, D'Ambrosio R, Dupont P. Kernel methods for heterogeneous feature selection. Neurocomputin. 2015. 169:187-195
4. Ferreira AJ, M.A.T. Figueiredo MAT. An unsupervised approach to feature discretization and selection. Pattern Recognition Letters. 2012. 45:3048–3060.
5. Liao B, Jiang Y, Liang W, Zhu W, Cai L, Cao Z. Gene selection using locality sensitive Laplacian score. IEEE/ACM Trans. Comput. Biol. Bioinform.2014. 11:1146–1156.
6. Mohammadi M, Noghabi HS, Hodtani GA, Mashhadi HR. Robust and stable gene selection via Maximum-Minimum Correntropy Criterion. Genomics. 2016. 107:83-87.
7. Tabakhi S, Moradi P, Akhlaghian F. An unsupervised feature selection algorithm based on ant colony optimization, Engineering Applications of Artificial Intelligence. 2014. 32:112-123.
8. Au WH, Chan KCC, Wong AKC, Wang Y. Attribute Clustering for Grouping, Selection, and Classification of Gene Expression Data, IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2005. 2(2): 83-101.
9. Wang D, Zhang H, Liu R, Lv W, Wang D. t-Test feature selection approach based on term frequency for text categorization, Pattern Recognition Letters. 2014. 45:1-10.
10. Piyush , Mundra A , Rajapakse JC. Gene and sample selection using T-score with sample selection. Journal of Biomedical Informatics.2016. 59:31–41.
11. Piyush K ,Mundra A, Rajapakse JC. Gene and sample selection using T-score with sample selection. Journal of Biomedical Informatics.2016. 59:31–41.
12. Anushaa M, Sathiaseelan JGR. Feature Selection using k-Means Genetic Algorithm for Multi-objective Optimization. Procedia Computer Science.2015. 57:1074 – 1080.
13. Vesanto J, Alhoniemi E. Clustering of the self-organizing map. IEEE Transactions on Neural Networks.2000. 11(3): 586-600.
14. Alshamlan HM, Badr GH, Alohali YA. Genetic Bee Colony (GBC) algorithm: A new gene selection method for microarray cancer classification. Computational Biology and Chemistry.2015. 56:49-60
15. Feng G, Jianhua Guoa, Jingc BY , Sunb T. Feature subset selection using Naive Bayes for text classification. Pattern Recognition Letters.2015. 65:109–115.
16. Jiang L, Li C, Wang S, Zhang L. Deep feature weighting for Naive Bayes and its application to text classification. Engineering Applications of Artificial Intelligence.2016. 52:26-39.
17. Novakovic J. The Impact of Feature Selection on the Accuracy of 1DwYHBayes Classifier. 18th Telecommunications forum TELFOR. 2010. pp. 23-25.
18. Zhang ML, José M. Peña JM, Robles V. Feature selection for multi-label Naive Bayes classification, Information Sciences. 2009. 179(19):3218–3229.
19. Loh WY. Classification and regression trees. WIREs Data Mining and Knowledge Discovery.2011. 1:14-23.
20. Akande KO, Owolabi TO, Olatunji SO.Investigating the effect of correlation-based feature selection on the performance of neural network in reservoir characterization. Journal of Natural Gas Science and Engineering.2015. 27(Part-1), pp. 98-108
21. B. Chandra B, Naresh Babu KV. Classification of gene expression data using Spiking Wavelet Radial Basis Neural Network. Expert Systems with Applications.2014. 41: 1326-1330.
22. Kundu PP, Mitra S. Multi-objective optimization of shared nearest neighbor similarity for feature selection. Applied Soft Computing. 2015. 37:751–762.
23. Cao J, Zhang L, Wang B, L Fi, Yang J. A fast gene selection method for mulit-cancer classification using multiple support vector description.Journal of Biomedical Informatics.2015. 53:381-389.
24. Liu D, Qiann H, Dai G, Zhang Z. An iterative SVM approach to feature selection and classification in high-dimensional datasets. Pattern Recognition.2013.46:2531–2537.
25. Thi HAL, Vo X, Dinh TP.Feature selection for linear SVMs under uncertain data: Robust optimization based on difference of convex functions algorithms.Neural Networks.2014. 59:36–50
26. Tong M, Liu KH, ChunguiXu, Ju W. An ensemble of SVM classifiers based on gene pairs. Computers in Biology and Medicine. 2013. 43:729–737.
27. StjepanOreski, GoranOreski. Genetic algorithm-based heuristic for feature selection in credit risk assessment. Expert Systems with Applications.2014. 41, pp. 2052–2064.
28. Pham HYN, Triantaphyllou E. A meta-heuristic approach for improving the accuracy in some classification algorithms.Computers & Operations Research.2011, 38:174–189.
29. Dadaneh BZ, Markid HY, Zakerolhosseini A. Unsupervised probabilistic feature selection using ant colony optimization. Expert Systems with Applications.2016.53:27-42
30. Maji P , Das C. Relevant and Significant Supervised Gene Clusters for Microarray Cancer Classification. IEEE Transactions on Nanobioscience.2012. 11 (2):161-168
31. Krejnik M, Klema J. Empirical Evidence of the Applicability of Functional Clustering through

Gene Expression Classification, IEEE Transactions on Computational Biology .Bioinformatics.2012, 9(3):788-798.

32. Ma PCH, Chan KCC. Incremental Fuzzy Mining of Gene Expression Data for Gene Function Prediction. IEEE Transactions on Biomedical Engineering. 2011.58(5):1246-1252

33. Wu GPK, Chan KCC, Wong AKC, Bin Wu .Unsupervised discovery of fuzzy patterns in gene expression data, IEEE International Conference on Bioinformatics and Biomedicine.2010.pp: 269-273

34. Fenga G, Guoa J, Jingc BY, Sunb T. Feature subset selection using Naive Bayes for text classification. Pattern Recognition Letters.2015. 65:109–115.

35. Kima C, HonglanLia, Shinb SY , Hwanga KB. An efficient and effective wrapper based on paired t-test for learning Naive Bayes classifiers from large-scale domains. Procedia Computer Science.2013.23:102 – 112.

36. Bermejo P, Gámez JA, Puerta JM. Speeding up incremental wrapper feature subset selection with Naive Bayes classifier. Knowledge-Based Systems.2014. 55:140–147

37. Chen KH, Wang KJ, Wang KM, Angelia MA. Applying particle swarm optimization-based decision tree classifier for cancer classification on gene expression data. Applied Soft Computing.2014. 24:773-780.

38. Tahir NM , Hussain A , Samad SA, Ishak KA. Feature Selection for Classification Using Decision Tree. 4th Student Conference on Research and Development.2006. pp.99–102.

39. Agrawal S, Agrawal J. Neural Network Techniques for Cancer Prediction: A Survey. Procedia Computer Science.2015. 60:769–774.

40. García-Laencina PJ, Sancho-Gómez JL , Figueiras-Vidal AR..Classifying patterns with missing values using Multi-Task Learning perceptrons. Expert Systems with Applications.2013.40:1333–1341.

41. Paul J, D'Ambrosio R, Dupont P. Kernel methods for heterogeneous feature selection. Neurocomputing.2015. 169:187-195.

42. Barkai UAN, Notterman DA,. Gish K, Ybarra S, Mack D, and Levine AJ. (1999) Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Arrays. Proc. Nat'l Academy of Sciences of the United States of Am.1999. 96 (12): 6745-6750.

43. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, E.S. Lander ES. Molecu.ar Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring, Science. 1999. 286:531-537

44. Liu L, Wong AKC, Wang Y. A Global Optimal Algorithm for Class-Dependent Discretization of Continuous Data. Intelligent Data Analysis.2004. 8 (2):151-170.

45. Li J, Wong L . Identifying Good Diagnostic Gene Groups from Gene Expression Profiles Using the Concept of Emerging Patterns. Bioinformatics.2002. 18 (5): pp. 725-734.

46. Mallick PK, Mishra D, Patnaik S , Shaw K. A Hybrid Approach for Simultaneous Gene Clustering and Gene Selection for Pattern Classification. Indian Journal of Science and Technology.2016. 9(21):pp.1-10

47. Li J, Wong L . Identifying Good Diagnostic Gene Groups from Gene Expression Profiles Using the Concept of Emerging Patterns. Bioinformatics.2002. 18 (10):1406- 1407.