Semi Supervised Ensemble Clustering Algorithm (GA Based) For High Dimensional Genomic Data

P.Krishnakumari [1] *K.Vivekanandan* [2]

Abstract

Clustering high-dimensional spaces are a difficult problem which is recurrent in many domains, for example in computational biology. Developing effective clustering methods for such domains are rare and also it is a challenging problem. This paper presents an efficient algorithm designed for high- dimensional gene data which combines the ideas of Linear Discriminant Analysis LDA based on PCA feature extraction along with K-Means algorithm to select the most discriminative subspace and it uses genetic algorithm for performing local optimization from the points that are globally optimal. The clustering process is thus integrated with the subspace selection process based on LDA and the data are then simultaneously clustered while the feature subspaces are selected .Then clustering instances are aggregated to generate final clusters based on agglomerative clustering. Also genetic algorithm is used to eliminate the problem of local optimality. Real datasets show that the proposed method outperforms existing methods for clustering high-dimensional genomic data in terms of accuracy and time.

[1]Senior Lecturer, Department of Computer Science, Sri Ramakrishna College of Arts and Science for Women , Coimbatore - 641044 , Tamilnadu, India. Email : kkjagadeesh@yahoo.com

[2]Reader, Bharathiar University, Coimbatore, Tamilnadu, India

Keywords : Gene expression, clustering, microarray analysis, K-Means clustering, Linear Discriminant Analysis, PCA, Genetic algorithm.

1. Introduction

Clustering in high-dimensional spaces is a difficult problem often referred as the "curse of dimensionality" for various application domains, such as information retrieval, computational biology, and image processing since the data dimension is usually very high for such applications. While various dimension reduction techniques have been proposed, there are two major types, feature transformation and feature selection [10][17]. Feature transformation methods project the original high dimensional space onto a lower dimensional space, while feature selection methods select a subset of meaningful dimensions from the original ones. The simplest approach of dimension reduction techniques includes principal component analysis (PCA) [9] [12] and random projections [6]. In these methods, dimension reduction is carried out as a preprocessing step and is decoupled with the clustering process. Once the subspace dimensions are selected, they remain fixed during the clustering process. An extension of this approach is the adaptive dimension reduction approach [8][14][15] where the subspace is adaptively adjusted and integrated with the clustering process. Subspace clustering algorithms that detect clusters in axis parallel to the projections of the original data set [13] are not able to capture local data correlations and find clusters of correlated objects since the principal axes of correlated

data are arbitrarily oriented. If we restrict the subspace to be linear combinations of original features, the subspace obtained in linear discriminant analysis (LDA) is perhaps the best subspace to do data clustering, because in LDA subspace, clusters are well separated. Bioinformatics and genomic sequence analysis, in particular is one of the hottest topics in modern science. The usefulness of statistical techniques in this field cannot be underestimated. The increasing use of DNA microarrays to generate large-scale datasets of gene expression has led to several important statistical and analytical challenges. Microarray experiments are being carried out in biological and medical researches to address a wide range of problems, including clustering of gene data [4] [3].

In traditional partitional based algorithms, problems due to initialization and local optima do arise. Also they find difficult to handle high dimensional data .Hence an algorithm is proposed to handle high dimensional data and also eliminates local optimality. The paper is organized as follows: Section 2 presents the related work, Section 3 presents theoretical problem definition for LDA based K-means clustering, Section 4 explains the proposed GA based learning framework by combining LDA ,K-means clustering and agglomerative clustering , Section 5 presents the experimental results, and finally Section 6 provides the conclusion.

2. Related Work

In 2001, LDA with K-Means is a very well developed theory with the growth of matrix-based approaches in machine learning [11][7][19][16][5]. LDA+K-Means algorithm reduces to the adaptive dimension reduction (ADM) algorithm [8] where only between-class scatter is optimized rather than the full LDA and adaptively modifies the subspace to fit the data distribution; here between-cluster scatter matrix is chosen explicitly. In 2004, LDA+K-Means algorithm reduces to the adaptive subspace iteration algorithm [14][15] where only the within-class scatter is optimized rather than the full LDA ;here between-cluster scatter matrix is chosen implicitly. In 2006, a matrix factorization [7] is proposed such that, after one matrix factor is eliminated; the two remaining matrix factors can be viewed as the projection directions in a LDA variant and cluster indicators respectively. They are solved in an alternative fashion using LDA and a soft-clustering similar to adaptive dimension reduction. In 2007 , LDA and K-Means clustering are simultaneously used as adaptive dimension reduction approach clustering [5] because they minimize the within-class scatter matrix and maximize the between-class scatter matrix and can be viewed as an unsupervised LDA. The proposed algorithm in this paper is developed based on this direction. In partitional clustering, problems due to initialization and local optima do arise. One way of approaching this challenge is to use stochastic optimization schemes such as genetic algorithms (Gas)[4].

The above algorithms adaptively modify the subspace to fit the data distribution when either the natural clusters in the data are close to spherical Gaussians or natural clusters are well separated. However there is a possibility of having suboptimal clusters as the clustered output directly depends on the selected dimensions for every run of the algorithm output. As a result there will be much deviation for the clustering results obtained and hence it does not guarantee for the good model for such data. Hence to modify the subspace adaptively to converge to the subspace where clusters are most separable and also to improve the clustering accuracy the proposed algorithm alternatively stores every obtained LDA

clustering instances in a similarity n x n matrix where n is the number of clustering instance generated. Based on the similarity matrix, an agglomerative clustering is finally applied to generate the final clustered result. The proposed algorithm combines LDA and K-Means in a simpler way and finally performs agglomerative clustering to improve the clustering accuracy for the high dimensional data. Thus the proposed algorithm alternatively combines selected dimensions and clustering and finally covers all the dimensions because for clinical applications coverage of all dimensions plays a major role. It is also integrated with genetic algorithm to provide global optimality.

3. Problem Definition

3.1 Linear Discriminant Analysis, C-classes: Derivation

Here (C-1) discriminant functions are used. The projection is from N-dimensional space onto (C-1) dimensions. The generalization of the within-class scatter matrix is

$$S_W = \sum_{i=1}^{c} S_i$$

$$\text{where } S_i = \sum_{x \in \omega_i} (x-\mu_i)(x-\mu_i)^T \text{ and } \mu_i = \frac{1}{N_i}\sum_{x \in \omega_i} x \qquad (1)$$

The generalization for the between-class scatter matrix is

$$S_B = \sum_{i=1}^{c} N_i(\mu_i-\mu)(\mu_i-\mu)^T$$

$$\text{where } \mu = \frac{1}{N}\sum_{\forall x} x = \frac{1}{N}\sum_{\forall i} N_i\mu \qquad (2)$$

where $S_T = S_B + S_W$ is called the total scatter matrix. For the (C-1) class problem, (C-1) projection vectors w_i are sought, which can be arranged by columns into a projection matrix $W=[w_1|w_2|\ldots|w_{C-1}]$ so that

$$y_i = w_i^T x \Rightarrow y = W^T x \qquad (3)$$

It can be shown that the optimal projection matrix W* is the one whose columns are the eigenvectors corresponding to the largest eigenvalues.

Since the projection is not scalar (it has C-1 dimensions), LDA produces the projection matrix W* that maximizes the following

$$J(W) = \frac{\left|\tilde{S}_B\right|}{\left|\tilde{S}_W\right|} = \frac{\left|W^T S_B W\right|}{\left|W^T S_W W\right|} \qquad (4)$$

3.2 K-Means Based LDA

The standard *K*-means clustering minimizes the clustering objective function

$$\min_H J_K,\ J_K = \sum_k \sum_{i \in C_k} \|x_i - m_k\|^2 \qquad (5)$$

where the matrix $H = \{0,1\}^{n \times K}$ is the cluster indicator: $H_{ik} = 1$ if x_i belongs to the *k*-th cluster, and $H_{ik} = 0$ otherwise. Tr *M* indicates the trace of matrix *M*. It is well-known that $S_t = S_w + S_b$. It is clear that the Kmeans clustering objective function is

$$J_K = \text{Tr}\, S_W = \text{Tr}\,(S_t - S_b) \qquad (6)$$

Therefore, K-Means clustering minimizes the within-class scatter matrix S_w, or maximizes the between-class scatter matrix S_b since Tr S_t is a constant. On the other

hand, given class labels as specified by the indicator matrix H, the LDA directions U are determined by

$$\max_{U} \mathrm{Tr} \frac{U^T S_b U}{U^T S_w U} \tag{7}$$

Thus LDA has very similar properties as *K*-means clustering: minimizing within-class scatter S_w and/or maximizing between-class scatter S_b. LDA is widely used to select the subspace (feature space) which has the maximal discriminant power. However, LDA is a supervised learning method which requires the class label for each data point before-hand. Since LDA and *K*-means clustering both minimizes S_w and maximize S_b, there should be ways to combine them into a single framework. In this paper, an algorithm is proposed to combine them into a single framework. K-Means clustering is used to generate class labels and use LDA to do subspace selection. The final results of this learning process are that the data are clustered while the feature subspaces are selected simultaneously and corresponding instances are stored and clustered using agglomerative clustering. LDA finds the most discriminative subspace in a unsupervised manner and optimizes the LDA objective function

$$\max_{U,H} \mathrm{Tr} \frac{U^T S_b U}{U^T S_w U} \tag{8}$$

Initially fix U and then obtain H as follows

$$\max_{H} \frac{\mathrm{Tr}\, U^T S_b U}{\mathrm{Tr}\, U^T S_w U} = \frac{\mathrm{Tr}\, U^T (S_t - S_w) U}{\mathrm{Tr}\, U^T S_w U} = \frac{\mathrm{Tr}\, U^T S_t U}{\mathrm{Tr}\, U^T S_w U} - 1.$$

Since $\mathrm{Tr}\, U^T S_t U$ is independent of H, this leads to

$$\min_{H} \mathrm{Tr}\, U^T S_w U = \mathrm{Tr} \sum_k \sum_{i \in C_k} U^T (\mathbf{x}_i - \mathbf{m}_k)(\mathbf{x}_i - \mathbf{m}_k)^T U = \sum_k \sum_{i \in C_k} \| U^T \mathbf{x}_i - U^T \mathbf{m}_k \|^2 \tag{9}$$

This is precisely the *K*-means clustering in the subspace $Y = U^T X$. Once H is calculated within and between cluster scatter matrices can be computed. U is given by d eigenvectors associated with the d largest eigenvalues of the between-cluster scatter matrix S_b

4. Proposed Algorithm

The proposed algorithm is based on LDA based adaptive dimension reduction approach that combines LDA and K-Means clustering, agglomerative clustering based on PCA feature extraction. As LDA and *K*-means clustering are optimizing the same objective function, i.e., they both minimize the within-class scatter matrix and maximize the between-class scatter matrix, it can be viewed as an unsupervised LDA. *K*-means clustering is used to generate class labels and use LDA to do subspace selection. The clustering process is thus integrated with the subspace selection process and the data are then simultaneously clustered while the feature subspaces are selected. Every clustering instance is stored in a n x n similarity matrix .Cluster membership is used as the bridge connecting the clusters discovered in the subspace and those defined in the full space. With this connection, clusters are discovered in the low dimensional subspace to avoid the curse of dimensionality and the results are aggregated to form an n x n similarity matrix where n is the number of instances. An agglomerative clustering is then applied to the matrix to produce final results.

4.1 Linear Discriminant Analysis (LDA)

Linear discriminant analysis (LDA) is the method used in statistics and machine learning to find the linear combination of features which best separate two or more

classes of objects or events. The resulting combination may be used as a linear classifier, or, more commonly, for dimensionality reduction before later classification.LDA is closely related to ANOVA (analysis of variance) and regression analysis, which also attempt to express one dependent variable as a linear combination of other features or measurements. In the other two methods however, the dependent variable is a numerical quantity, while for LDA it is a categorical variable (i.e. the class label).LDA is also closely related to principal component analysis (PCA) and factor analysis in that both look for linear combinations of variables which best explain the data. LDA explicitly attempts to model the difference between the classes of data. PCA on the other hand does not take into account any difference in class, and factor analysis builds the feature combinations based on differences rather than similarities. Discriminant analysis is also different from factor analysis in that it is not an interdependence technique: a distinction between independent variables and dependent variables (also called criterion variables) must be made.

4.2 Aggregating Multiple Clustering Results

The clustering results are aggregated into a matrix that measures the similarity between each pair of data points. Then an agglomerative clustering algorithm is applied to produce the final clusters. For each subspace selection and clustering process, if θ represents a model then for each data point i, the soft clustering results $P(l \mid i, \theta)$, $l = 1,\ldots,k$ are given, representing the probability that the point belongs to each cluster under the model θ. P_{ij}^{θ} can be defined [18] as the probability of data point i and j belonging to the same cluster under model θ and it can be calculated as :

$$P_{ij}^{\theta} = \sum_{l=1}^{k} P(l|i,\theta) \times P(l|j,\theta) \qquad (10)$$

To aggregate multiple clustering results, the values of P_{ij}^{θ}'s are averaged across n runs to obtain P_{ij}, an estimate of the "probability that data point i and j belong to the same cluster". This forms a similarity matrix. This is tested by performing ten runs of the algorithm on the synthetic data set and separated the aggregated P_{ij} values into two groups based on if data point i and j are from the same cluster. Pij values are large when data point i and j are from the same natural cluster and small otherwise.

4.3 The Agglomerative Algorithm

To produce the final clusters from the aggregated similarity matrix P, an agglomerative clustering is applied as follows

Algorithm:

Inputs: P is a n x n similarity matrix, k is a desired number of clusters.

Output: a partition of n points into k clusters.

Procedure: An Agglomerative clustering Algorithm

$l = n$.

For i = 1 to n

Let $c_i = \{x_i\}$ for $i = l,\ldots,n$

Repeat

Find the most **similar** pair of clusters based on P, say c_i and c_j.

Merge c_i and c_j and decrement l by one

Until $l <= k$

The similarity between two clusters is given as follows

$$sim(c_i, c_j) = \min_{x_i \in c_i, x_j \in c_j} P_{ij} \qquad (11)$$

When two points have very small similarity value (i.e., small possibility of belonging together according to P) the algorithm will not group them together.

4.4. The Proposed LDA Method

A unique feature in this approach is switching between the subspace (for clustering) and the original space (for LDA). The cluster indicator H enables us to uniquely connect the the spaces. With this connection, clusters are discovered in the low dimensional subspace to avoid the curse of dimensionality and are adaptively re-adjusted for global optimality.

Algorithm Steps:

Step 1 : Set dimension d to be K-1 where K is number of clusters

Step 2 : Generate class label performing K-means clustering based on initial U obtained from PCA and compute H

Step 3 : Store the clustering instance

Step 4 : From H, derive next LDA subspace U and perform k-means clustering

Step 5 : Repeat steps 3, 4, 5 until convergence

Step 6 : Construct the n x n similarity matrix where 'n 'is the number of instances

Step 7 : Perform Agglomerative clustering to produce final clusters

As $Y = U^T X$ is proportional to P_{ij} aggregation , we can write $U^T X = K P_{ij}$. Since K does not depend on class labels, this is effectively close to equation (5).

The proposed partitional based ensemble method is deemed to be the best at that point in the algorithm, but may not be the best globally when all information is considered and hence it is integrated with genetic algorithm to find optimal or near optimal solutions on complex, large spaces of possible solutions. In GAs, biologically inspired operators like crossover and mutation are applied to yield a new generation of strings [4] based on fitness. The process of selection, crossover and mutation continues for a fixed number of generations or till the termination condition is satisfied.

4.5 GA Based Ensemble Algorithm

```
Algorithm
{
Randomly generate K cluster centers from initial
population;
Perform proposed LDA method;
While (Not termination condition) do
{
Evaluate fitness ( );
Select ( );
Apply crossover ( );
Apply mutation ( );
Perform proposed LDA ensemble clustering;
}
Show Clusters Based on the Best Cluster Centers
}
```

Fitness is determined based on the Euclidean measure. Here single point crossover with a crossover probability of 0.9 is chosen. Mutation rate is 0.1. Best individuals are selected based on roulette wheel selection. The computational complexity of the GA based algorithm is $O(gpnktd) + O(d^2nt) + O(p1^2)$ for K-Means clustering combining LDA and aggregating clustering instances where d is the dimension of data, n is the number of data points, t is the number of iterations, k is the number of clusters, p1 is the number of instances, g is number of generations, p is the population.

5. Data Sets

High density DNA microarray technology can simultaneously monitor the expression level of thousands of genes which determines different pathological states of the same tissue drawn from different patients

[1][20].The proposed algorithm is implemented in MATLAB to analyze well known data sets. Two gene data sets with relatively high dimensions are chosen from the NCBI database. First data set contains drosophila melanogaster genes. The data consists of 10000 genes, of which only 8175 genes are identified to respond significantly. The second data set is the real AML-ALL leukemia data set. AML-ALL is a gene sample data set which consists of 38 bone marrow samples over 7129 probes from 6817 human genes.

6. Experiment Results

The proposed algorithm is implemented in MATLAB. The fig1 shows the LDA projection in the first step.

The proposed GA based algorithm is implemented for the two real data sets and the results are tabulated in the tables 1, 2. Table 1 shows the performance of the proposed GA based algorithm in terms of time and accuracy for the two data sets. It is observed that the proposed algorithm takes less time at higher dimensions. It is observed from the Table 2 & 3 that the mean, standard deviation and coefficient of variance for GA based ensemble are less than GA based k-means with PCA and also the GA based ensemble is more consistent because the coefficient of variance is low.

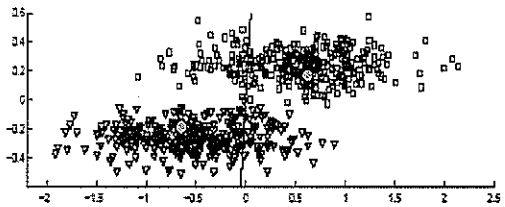

Figure 1 : shows the LDA projection in the first step

Table 1: Performance of the Proposed GA Based Ensemble Algorithm In Terms Of Time and Accuracy For Drosophila Melanogaster Gene Data Set 1 And Leukemia Data Set 2

No of Dimension		GA Based Ensemble (data set 1)	GA Based K-Means (data set 1)	GA Based Ensemble (data set 2)	GA Based K-Means (data set 2)
Time (Sec)	100	8.12	8.91	9.02	10.70
	200	8.77	9.51	9.17	11.20
	300	9.33	10.84	10.23	13.37
	400	10.18	12.92	10.91	13.92
	500	11.11	13.93	11.15	14.33
Accuracy	100	0.821	0.619	0.661	0.410
	200	0.732	0.596	0.651	0.444
	300	0.646	0.422	0.637	0.531
	400	0.716	0.489	0.799	0.519
	500	0.677	0.423	0.777	0.513

Table 2: Results Of The GA Based Ensemble Clustering In Terms Of Time For Data Set 1 & 2

Results	GA based Ensemble (data set 1)	GA based K-Means (data set 1)	GA based Ensemble (data set 2)	GA based K-means (data set 2)
Mean	2281.11	2289.29	2432.10	2438.29
Standard Deviation	2.63	2.71	3.21	3.33
Co - efficient of Variance	0.1144	0.1178	0.1319	0.1368

t value =6.8503 *p* value = 0.000045 *t* value =4.232 *p* value = 0.001738

Table 3: Results of The GA Based Ensemble Clustering In Terms Of Accuracy For Data Set 1 & 2

Results	GA based Ensemble (data set 1)	GA based K-Means (data set 1)	GA based Ensemble (data set 2)	GA based K-means (data set 2)
Mean	2487.20	2480.12	2391.10	2381.10
Standard Deviation	2.96	3.19	3.03	3.12
Co - efficient of Variance	0.1190	0.1286	0.1267	0.1310

t value =5.14 *p* value = 0.000438 *t* value =7.27 *p* value = 0.000027

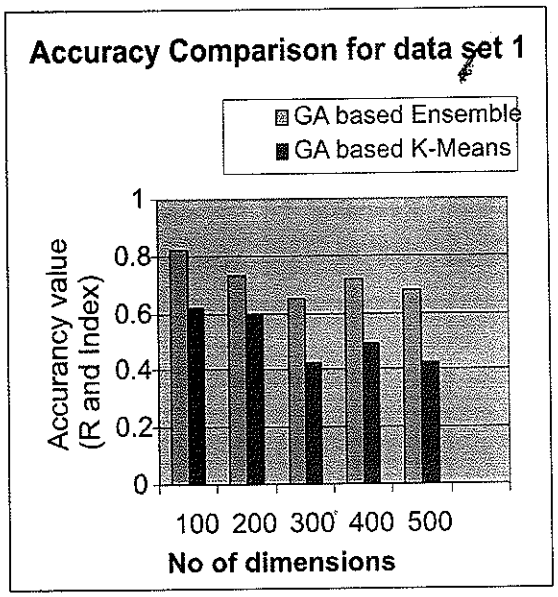

Figure 2: Accuracy Comparison For Data Set 1

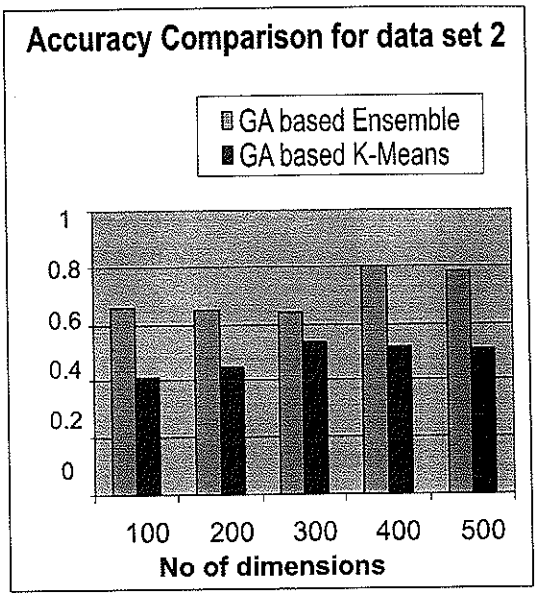

Figure 3: Accuracy Comparison for Data Set 2

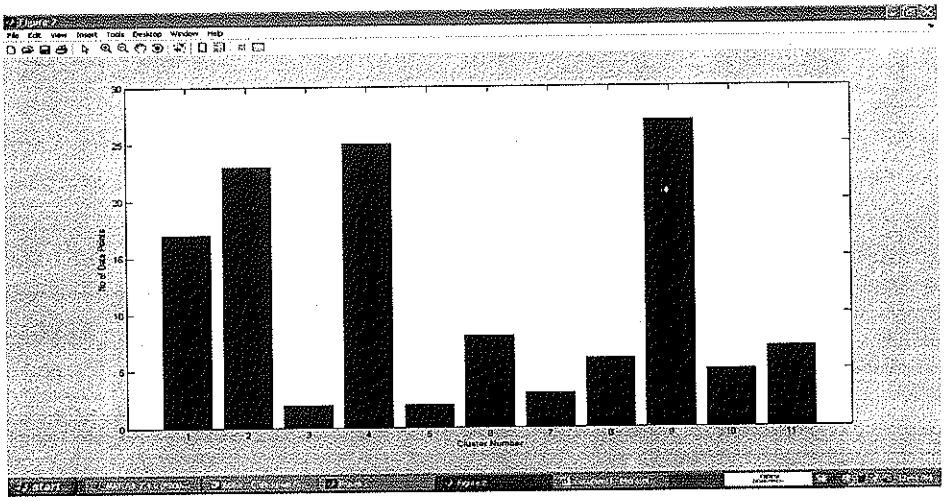

Figure 4: Sample Output Showing the Clusters Formed For GA Based Ensemble for K=11

The accuracy is calculated using Rand index. Table 1 show that the accuracy for the proposed method is high. From the Table 2 the two-tailed P value equals 0.000045 and 0.001738 for the two data sets. From the Table 3, the two-tailed P value equals 0.000438 and 0.000027 for the two data sets. By conventional criteria, these differences are considered to be statistically significant. It is observed by *t*-test analysis that the *p* value is less than 0.05 and hence there exists significant correlation between the methods. It is concluded that there is significant difference between proposed method and K-Means with PCA with respect to accuracy and time.

7. Conclusion And Future Enhancement

In this work, a new ensemble algorithm is proposed that handles all the dimensions efficiently. The proposed clustering algorithm based on GA has been implemented and tested successfully using MATLAB on windows operating systems. The results show that GA based

clustering provides global solution with significant results. The time complexity of the proposed algorithm is relatively less for high dimensional data and produces better accuracies. Since the proposed algorithm helps in reduction of dimension it considerably reduces the space. In future parallel computing techniques can be applied to increase the speed of GA.

References

1. Alizadeh A.A , Eisen M . B et al , *"Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling"*, Nature, 403:503–511, 2000.
2. 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"*, *Proc Natl Acad Sci USA*, 96:6745–6750. doi:10.1073/pnas.96.12.6745. [PubMed], 1999.
3. Cadez .I, Gaffney . S and Smyth .P Technical Report UCI-ICS-00-09, University of California, Irvine, 2000 .
4. Chakraborty .A and Maka .H, *"Biclustering of Gene Expression Data Using Genetic Algorithm"*, Proceedings of the IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, Vol. 14, No. 15, PP.1 – 8, 2005.
5. Chris Ding and Tao Li , *"Adaptive Dimension Reduction Using Discriminant Analysis and K-means Clustering"*, Proceedings of the 24th International Conference on Machine Learning, Corvallis, OR, 2007.
6. Dasgupta. S, *"Experiments with random projection: Uncertainty in Artificial Intelligence"*, Proceedings of the Sixteenth Conference (UAI-2000), PP. 143-151, Morgan Kaufmann 2000.
7. De la Torre .F and Kanade .T, *"Discriminative cluster analysis"*, Proc. Int'l Conf. Machine Learning , 2006.
8. Ding C, He X, Zha .H and Simon .H, *"Adaptive dimension reduction for clustering high dimensional data"*, Proc. IEEE Int'l Conf. Data Mining, 2002.
9. Duda .R .O, Hart .P .E & Stork .D.G, *"Pattern classification"*, 2nd edition, Wiley, 2000.
10. Fodor I K, *"A survey of Dimension Reduction Techniques"*, LLNL Technical Report, UCRL ID-148494, URL: http://www.llnl.gov/CASC/sapphire/pubs.html, 2002.
11. Hastie .T, Tibshirani .R and Friedman. J, *"Elements of statistical learning"*, Springer Verlag, 2001.
12. Jolliffe .I , *"Principal component analysis"*, Springer, 2nd edition, 2002.
13. Kailing .K, Kriegel and Kroger .P, *"Density connected subspace clustering for high-dimensional data"* , Proc. 4th SIAM Int. Conf. on Data Mining , Florida, 2004.
14. Li .T and Ma .S , *"IFD: Iterative feature and data clustering"*, Pro. SIAM Int'l conf. on Data Mining, PP. 472–476, 2004.
15. Li .T, Ma .S and Ogihara .M, *"Document clustering via adaptive subspace iteration"*, Proc. conf. Research and development in IR (SIRGIR) , PP. 218–225, 2004.

16. Park.H and Howland .P, *"Generalizing discriminant analysis using the generalized singular value decomposition"*, IEEE. Trans. on Pattern Analysis and Machine Intelligence, 26, 995 – 1006, 2004.

17. Parsons.L et al, *"Subspace Clustering for High Dimensional Data: a Review"*,ACM SIGKDD Explorations Newsletter, 6(1), 90 – 105,2004.

18. Xiaoli Zhang Fern, Carla E Brodley, *"Random Projection for High Dimensional Data Clustering: A Cluster Ensemble Approach"*, Proceedings of the Twentieth International Conference on Machine Learning , Washington , 2003.

19. Ye .J and Xiong .T , *"Null space versus orthogonal linear discriminant analysis"*, Proc. Int'l Conf. Machine Learning, 2006.

20. Yeoh .E .J, Ross .M .E, Shurtleff .S A, Williams .W .K, Patel .D, Mahfouz. R, Behm .F .G, Raimondi .S .C, Relling. M .V, Patel .A, Cheng. C, Campana. D, Wilkins .D, Zhou .X, Li. J, Liu .H, Pui. C. H, Evans .W .E, Naeve .C, Wong .L, Downing.J.R, *"Pediatric Lymphoblastic Leukemia by Gene Expression Profiling"*, Cancer Cell, *1*:133–143, doi: 10.1016/S1535-6108(02)00032-6, 2002. [PubMed]

Author's Biography

Dr. K. Vivekanandan received the Ph.D. degree in Computer science from Bharathiar University, India. He is currently Reader in Bharathiar University, India. He has a total teaching experience of 21 years. He has published 22 papers in international and national journals .He has produced 5 Ph D's and 11 MPhil scholars in computer science. His research interests include data mining and knowledge discovery, and Management Information System. .

P. Krishnakumari received the MPhil degree in Computer science from Bharathiar University, India. She is currently pursuing her Ph.D in computer science and has a total 13 years of teaching experience. At present she is a senior lecturer in the Department of computer science of Sri Ramakrishna college of Arts and Science for women, Coimbatore. She has presented 9 research papers in international and national conferences and published 6 papers in international and national journals. Her research interests include data mining and knowledge discovery, genetic algorithms, Bioinformatics, image compression, networking.