Proceedings of the 2007 IEEE Symposium on Computational
Intelligence in Bioinformatics and Computational Biology (CIBCB 2007)

Two-way Clustering using Fuzzy ASI for Knowledge Discovery in Microarrays

J. Shaik and M. Yeasin
Computer Vision, Pattern and Image Analysis Lab (www.cvpia.org)
Electrical and Computer Engineering,
University of Memphis,
Memphis, TN- 38152

Abstract- **This paper presents two-way clustering of microarray data using fuzzy adaptive subspace iteration (ASI) based algorithm for knowledge discovery in microarrays. It is widely believed that each gene is involved in more than one cellular function or biological process. The proposed fuzzy ASI assigns a relevance value to each gene associated with each cluster. These functional categories are ranked based on their potential in providing maximal separation between the two tissues classes; which is an indication of differentially expressed genes (DEGs). Empirical analyses on simulated, 100 artificial microarray datasets are used to quantify the results obtained using the fuzzy-ASI algorithm. Further analyses on different microarray cancer datasets revealed several important genes that are relevant with various cancers.**

Keywords- Fuzzy Clustering, Two-way clustering, Knowledge discovery in microarrays, visualization.

I. Introduction

Microarrays measure expression of thousands of genes under some experimental conditions. These conditions are normally labeled on the basis of some external information such as, clinical identification of tissue samples or expression of genes with respect to time [1]. Genes relevant to the pathology under investigation are expected to be up or down regulated between healthy and diseased tissues. The ranking of the genes depends on the feature selection techniques. The genes are generally ranked based on two criteria, i) individual differential expression of a gene between two tissue cases, or ii) co-expressed genes offering high discrimination between two tissue cases. The co-expression depends on the complex interactions between the genes. Both of these criteria require filtering of irrelevant genes (feature selection) for further processing. The feature selection is an important problem because only a subset of genes may be responsible for a biological process (for example, formation of tumor). The rest of the genes form noise and mask the underlying message. Most of the techniques based on co-expression assign each gene to a single cluster. It is widely known that each gene might be involved in more than one biological process. The allocation of each gene to a single cluster does not ensure this characteristic of the genes. Hence it is necessary to design a clustering algorithm which assigns a gene to multiple functionally relevant clusters based on their membership values to each cluster. The fuzzy clustering process addresses the process of one gene involving in more than one cellular process.

The proposed Fuzzy-ASI clustering algorithm as shown in Fig. 1 employs a progressive two-way clustering of the expression data to functionally classify the genes and to find DEGs [2-8]. The two-way clustering method employs one-way clustering in both dimensions. The clustering of one dimension is dependent on the clustering of the other dimension. Although variations of this method exist, most algorithms involve all the samples in the clustering process. One can possibly argue that samples that are noisy may result in false alarms. This problem can be addressed by employing an Adaptive Subspace Iteration (ASI) algorithm which performs all its calculations in Eigen-space, minimizing such possibilities. Further noise may be eliminated by employing fuzzy clustering [9]. The noise points are represented by low membership values with in a cluster. By employing a proper threshold the algorithm can be made robust to the noise. The two-way clustering process as shown in Fig. 2 provides insight into genes having similar functions. The functionality of unknown genes is found in the process of grouping of the genes. The samples are clustered using the gene clusters formed ignoring the tissue labels initially and later the label information is used as ground truth to rank the clusters.

The clustering methods which have been used rigorously include hierarchical methods [10-12], self organizing maps [13] and k-means clustering [14]. All these algorithms come under heuristic based approaches. In hierarchical clustering, the relationship among the variables (genes) is shown by a tree which depicts the similarity or dissimilarity among the groups of features. The advantage of such a method is that one may focus on more interesting details but, these methods lack robustness. The unsupervised projection-based method, for example, the self organizing map (SOM), provides low-dimensional representation of an input data space. This method is easy to implement but for large datasets this algorithm is not favorable and is computationally intensive. The K-means algorithm requires apriori knowledge about the data to initiate the algorithm and is very sensitive to the initialization. Further, choosing an algorithm that suits the data under consideration is problematic. Also most of the algorithms have similar performance which makes the choice

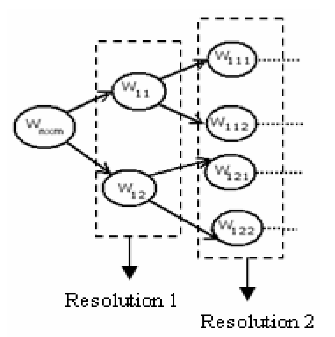


Fig 1: A framework for progressive clustering.

1-4244-0710-9/07/$20.00 ©2007 IEEE

much more difficult. It is also well known that all the heuristic algorithms have one inherent problem of determination of number of clusters.

A number of fuzzy based clustering algorithms are employed for microarray data. The most common algorithm used is the Fuzzy-C-means algorithm [15]. Many variations of this algorithm exist although most of them focus mainly on the fuzzification factor [16]. The subspace-based methods are relatively faster and computationally efficient and also produce good clustering of very high dimensional data. The Fuzzy-ASI algorithm [17] employed in this paper is a variant of the nested subset method (hard clustering method) and provides best possible feature selection for clustering [18, 19]. In this paper, Davies-Bouldin index is used to measure the quality of clusters for varying number of classes [20]. The number which offers highest quality clusters is considered optimum. Often this is estimated by plotting the quality indices vs number of clusters. The number of clusters corresponding to the knee of such a plot is a good estimate of the number of clusters in the data.

The Fuzzy-ASI based algorithm is expected to play an important role in data mining applications as it uses the synergy between dimensionality reduction and soft- clustering. The subspace-based algorithm is unique in the way it performs clustering and also provides interpretation of the groups formed in the process of clustering [17, 21]. The ASI algorithm simultaneously performs: a) dimensionality reduction, b) identification of subspace structure associated with each cluster and c) updates the memberships based on the subspace structures. Similarity between the data points is found in the Eigen-space and hence, the clustering of the data is robust against arrangement of samples. The ASI algorithm is also very robust against initialization conditions as the optimization procedure involved here iterates over the fore mentioned steps until a local minimum is reached. Additional features include automatic determination of the importance of features in the formation of clusters, dynamic allocation of new data points to clusters without repeating the whole process, automatic estimation of centroids of each cluster, ease of implementation, less computational intensive and ability to handle high dimensional data.

The two-way clustering mechanism identifies genes with a correlated level of expressions and these gene groups are ranked based on their potential to discriminate the known tissue cases. The ranking of the cluster depends on the ability of the co-expressed genes in the cluster to offer high discrimination between the samples. The clusters which offer high discrimination are highly ranked and vice versa. The high ranked clusters are potential candidates for finding the differentially expressed genes (DEGs). The performance of the two-way clustering technique heavily depends on the framework used for computing. For example, Tang et al. reported an inter-related clustering format [8] based on an iterative process which is very tedious and uses heuristics to define the number of clusters. McLachlan et al. assume a model of distribution to cluster the genes [22]. This process is limited by the unique characteristics of the microarray data which has limited samples. Getz et al. proposed a procedure called coupled two way clustering [6] by iteratively applying one way clustering with in the subgroups of genes and tissue clusters from the previous iteration. This method leads to too many iterations resulting in a tedious search as groups become smaller and smaller [23].

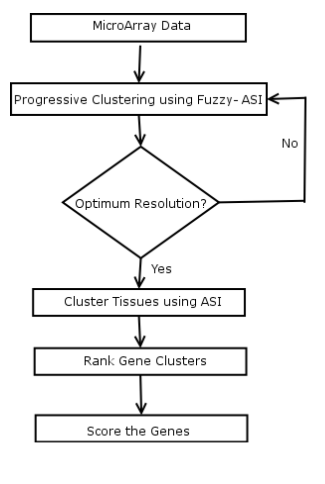


Fig. 2. Flowchart of two-way clustering using the ASI algorithm for functionally classifying genes and finding the differentially expressed genes.

The two-way clustering algorithm as shown in Fig. 2 on the other hand analyzes the gene groups into all possible resolutions as shown in Fig. 1. The ability of clusters to differentiate different tissue cases is then studied at different resolutions. The clusters which offer meaningful sample clustering are analyzed further. The tissue groups obtained using each gene group is compared with the actual class label (for example, tumor or normal) of the tissues. The gene clusters providing tissue clusters with high consistency with respect to the label information are highly ranked and vice versa. Please note that the genes are ranked here group-wise but not individually. Also, no assumption about the distribution of the data is made. This process extracts co-expressed genes based on maximal separation between two tissue cases.

This paper also uses an automated 3D star coordinate based projection (3D SCP) for projection of tissues using DEGs formed [24]. A good analysis of performance of all these techniques may be found in [18, 24]. Some of the advantages of 3D SCP include (not limited to): i) Dynamic projection of data points, ii) Multiple views of visualization, iii) Relative ease of interpretation, iv) Ease of implementation, v) Visual clusters and vi) No human intervention. In this paper, PCA and automated 3D SCP are employed for visualization and validating the DEGs obtained using the unified framework.

II. Mathematical background

The Fuzzy-ASI is an iterative method to cluster the data. It involves an optimization process that iteratively identifies the subspace structure. A subspace structure is a linear combination of original feature space. The weights determine the samples that are important in formation of the cluster and also provide data reduction. Let us consider that W_{nxm} is the dataset where the symbols $'m','n'$ represents the number of samples and the number of genes, respectively. Also assume that there are $'k'$ number of clusters; D_{nxk} is the partition matrix; F_{mxk} is the subspace structure associated with each cluster. The columns of the F matrix contain the weights of the samples. Hence, $(WF)_{nxk}$ represents the projection of the data onto the subspace. Let $'S'$ be the projection of centroid of each clusters onto the subspaces defined by the matrix $'F'$. The relationship between the $'S'$, 'F' and 'D' is given by:

$$S = (D^T * D)^{-1} * D^T * W * F. \tag{1}$$

The optimization function $'O'$ is given by:

$$O = \frac{1}{2} \| WF - DS \|^2_F . \tag{2}$$

Fuzzy ASI Algorithm

```
Begin clustering
Step 1: Begin Initialization
        Initialize '$D$' with random values such that each row adds up to 1.
        Initialize 'F' with Random values;
End Initialization
Step 2: Compute 'S' using equation (1);
Step 3: Compute '$O_0$' using equation (2);
Step 4: Perform optimization;
        Begin Optimization
        While ($O_1 < O_0$)        //Continue as long as the optimization value decreases
Step 4-1: Update '$D$' given by the formulae in equation (4)
      for i = 1 to n
        $P(j) = \|(WF) - S\|$;    j=1...k    (3)
        Step 4-2: Membership Assignment
        $D(i,j) = P(i,j) / \sum_j p(i,j)$;    j=1...k    (4)
        end for
Step 4-3: Update '$F$' given by the formula in equation (5);
        $F = E((W^T(I_n D(D^T D)^{-1} D^T - I_n)W)_{1:k})$; (5)
        Step 4-4: Compute Step 2;
        Step 4-5: Compute '$O_1$' using equation (2);
Step 4-6: If ($O_1 < O_0$) ;    //Check for the    terminating condition//.
   $O_0 = O_1$;
        End optimization
End Clustering
```

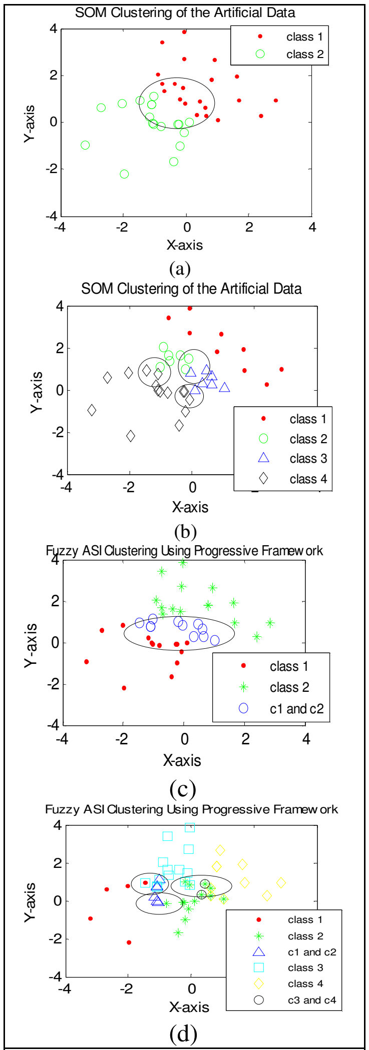
In Eq. 3, P is a similarity measure between the vectors. In Eq. 5, 'E' represents eigen vectors of the enclosed equation and $1:k$ represents first $'k'$ eigen vectors corresponding to the highest eigen values. The output of the algorithm is 'D' and 'F'. Here, 'D' offers the cluster memberships and 'F' offers the weights of the samples forming the clusters defined by the matrix 'D'. The elements of the matrix $(D^T D)$ contain the size of the clusters. Using the fuzzy-ASI clustering, genes may belong to multiple clusters with different memberships. Based on the membership, biological significance and hence the relevance of the gene with a cluster may be estimated. If the membership is low, the relevance is low and hence such genes form noise and must be pruned from the cluster. A choice of an appropriate threshold is hence necessary. The progressive framework besides providing partitions at user defined resolution, also provide an interesting choice for the threshold. At each resolution, since each of the clusters has to be divided into two partitions, the genes with memberships near 0.5 may be assigned to both the clusters and the ones with high memberships may be assigned to the respective clusters. The proper choice of number of levels for the progressive framework is obtained using cluster validation metrics. Davies-Bouldin index [20] is used to estimate the quality of clusters at every iteration, lower the value, better is the quality of clusters. The resolution level and hence the number of clusters which provide highest quality of clusters is chosen. A plot of Davies-Bouldin index value at each resolution is plotted. The elbow in the plot as shown in Fig. 5 is used as an indication of proper choice of resolution at which the progressive framework can be terminated. More of this will be discussed in the Section III. The differentially expressed genes are obtained using two-way clustering of microarray data using Fuzzy-ASI for $n = 25$ different iterations and the number of times a gene is found to be differentially expressed is listed. A gene may be listed more than once in each iteration because each gene might be a part of multiple clusters that are functionally related to the genes. Such genes are highly scored. Each gene is given a confidence value by summing up the score for 25 iterations and then normalized to obtain a value between $(0,1]$. Genes with high confidence value are chosen to be differentially expressed and further studied. The biological aspects of this list are also thoroughly studied as shown in Section III.

III. EMPIRICAL ANALYSES

Empirical analyses on both simulated and real microarray data sets were conducted to illustrate the efficacy of the proposed Fuzzy-ASI based soft clustering method. The main objective is to compare the performances of both soft and crisp clustering algorithms on same data set. In particular, the crisp clustering based methods such as SOM, ASI-based crisp clustering and the proposed Fuzzy-ASI based clustering are compared using 100 artificially generated microarray datasets. Further analyses were performed using a number of real microarray cancer datasets viz. Gastric cancer, colon cancer and leukemia. The empirical analysis shows that Fuzzy-ASI performed well in uncovering potential genes that may be involved in pathogenesis using microarray data.

Case Study 1: Simulated dataset 1

The artificial dataset is simulated to have two non-separable classes along the principal diagonal. The first class is generated by using the relation $x = y = t$. Where $'t'$ takes on 20 values equally spaced between -1 and 1. A Gaussian noise of mean 0 and standard deviation 1 is added independently to x and y. The second class is generated similarly but by adding a constant 1. The noise added is independent to the noise added to class 1. The data is clustered using Fuzzy-ASI and compared to results using SOM which provides hard clustering. The SOM clustering of the data is shown in the Fig. 3 (a). The points near the boundary of two clusters have high relevance with both the clusters. Such relevancy however may not be captured using hard clustering. Fuzzy-ASI using progressive framework at resolution 1 on the other hand as shown by ellipse in Fig. 3(c) assigns those points to both clusters with relevant memberships. The ‘c1 and c2’ in the legend of Fig. 3(c) represent points assigned to both the clusters with relevant memberships. Using the Davies-Bouldin index, the optimum number of clusters is estimated to be 4 for SOM and the optimum resolution was estimated to be 2 for Fuzzy-ASI using progressive framework. Fig. 3(b) shows the 4 clusters obtained using SOM.



(a)

(b)

(c)

(d)

Fig. 3. Clustering of simulated dataset1 a) Two clusters using SOM, b) four clusters using SOM c) Fuzzy-ASI at progressive clustering level 1 and d) Fuzzy-ASI at progressive clustering level 2.

As shown in Fig. 3(b), the boundary of separation is not clear. There are multiple points as shown by the circles which have relevance with adjacent clusters. Those points as shown in Fig. 3(d) have been assigned to multiple clusters with appropriate memberships. The ‘c1 and c2’ & ‘c3 and c4’ in the legend of Fig. 3 (d) represent points common to ‘class1 and class 2’ & ‘class 3 and class 4’, respectively. The overlapping points as shown by the circles in Fig. 3(d) represent points belonging to multiple clusters with appropriate memberships. As shown in Figs. 3(c) and (d), the progressive framework provides relationships among the clusters at user defined resolution.

Case Study II: Artificial microarray datasets

It was established in [25] that microarray datasets follow lognormal distribution. The artificial microarray datasets used in this paper are generated based on a hierarchical model. The data in each of the classes are drawn from normal distributions. The prior distributions of means are normally distributed and variances follow a gamma distribution. Means are chosen in such a way that data from different cases have small differential expression for DEGs and marginal or no differential expression for non differential expressed genes (NDEGs). Unequal variances are used for both cases for differentially expressed genes as proposed in [25, 26].

The performance of three clustering algorithms viz. Fuzzy-ASI, Hard-ASI and SOM following two-way clustering are studied using 100 artificially generated microarray datasets. It has previously been established that Hard-ASI based algorithm performed better for knowledge discovery in microarray data when compared to well known ranking and clustering algorithms [19, 27-29]. In this paper, the performance of Fuzzy-ASI is compared with that of Hard-ASI and SOM. As shown in Fig. 4, Fuzzy-ASI performed better in finding DEGs from microarray data. The true positive fraction (TPF) and false positive fraction (FPF) are calculated by finding number of genes found to be differentially expressed using the algorithm and indeed are differentially expressed when compared with ground truth and genes found to be differentially expressed and are not differentially expressed when compared with ground truth. For Fuzzy-ASI & Hard-ASI, the resolution level, and for SOM the number of clusters is estimated using Davies-Bouldin indices. Fig. 5 shows the index values calculated using Davies-Bouldin index for Fuzzy-ASI algorithm on one artificially generated microarray dataset. The elbow point in Fig. 5 shows that resolution level 9 is optimum for this dataset and hence is the terminating point for the progressive framework shown in Fig. 1.

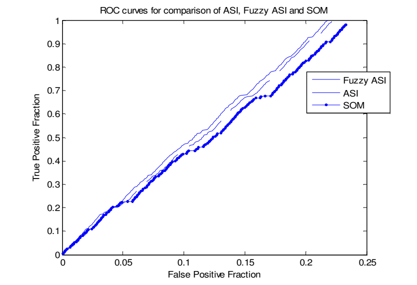


Fig. 4. Roc curve showing performance of Fuzzy-ASI with respect to ASI and SOM

Real Microarray Datasets

The relatively better performance of Fuzzy-ASI in knowledge discovery is established using artificial microarray datasets. The Fuzzy-ASI algorithm is now applied on real cancer datasets to functionally classify the genes and to find differentially expressed genes. Please note that the differential expression of the genes is not found by ranking the genes independently but by ranking the gene clusters based on their ability to differentiate different tissue cases.

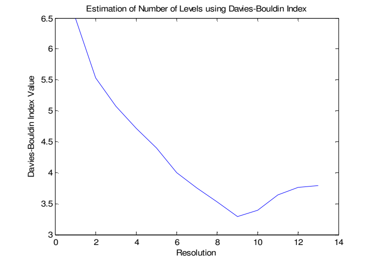


Fig. 5. Estimation of number of levels required for clustering of artificial microarray data using Davies-Bouldin Index

Case study III: Gastric Cancer Dataset

The objective of this empirical analysis is to identify genes distinguishing primary gastric cancers and metastatic gastric cancers from neoplastic gastric cancers which are otherwise morphologically indistinguishable. Approximately 30300 genes are used to study expression patterns of 90 primary gastric cancers, 14 metastatic gastric cancers and 22 neoplastic gastric cancers. The preprocessing steps mentioned in [30] are used resulting in 5200 genes for further study. The two-way clustering using Fuzzy-ASI is applied to find DEGs between neoplastic gastric cancers and other gastric cancers. The co-expressed genes which are also consistently differentially expressed with high confidence value are:

Metallothionein 1G (confidence 100%), Metallothionein 1F(100%), Metallothionein 1F functional DNA(100%), RecQ protein like 5 (100%), MARCKS-like 1 (100%), Early estrogen induced gene 1 protein (96%), Secretoglobin family 2A member 1(96%), Somatostatin (96%), transmembrane family, member 2 (96%), transmembrane family member 2 (84%), Beta 1,4-N-acetylgalactosaminyl transferase-transferase-III (92%), Acyl-Coenzyme A oxidase 1, palmitoyl (84%), diacylglycerol kinase delta(68%), signal induced proliferation associated 1 like 3 (75%), dermatopontin (75%), sulfotransferase family, cytosolic 1C member 1 (75%), protein disulfide isomerase family A, member 2 (73%), glutathione peroxidase 3(plasma) (77%) and cysteine and glycine rich protein 3 (77%).

Analysis indicates that several potential genes responsible for the disease are identified by the algorithm. These genes include (but not limited to) 4 Metallothionein genes, 2 tumor proteins P53, 4 Gastric cancer related proteins, 4 trasmembrane proteins etc.

Fig. 6 shows the projection of different tissues projected using DEGs as features using 3D SCP. As

shown in Fig. 6, different tissues were clearly separated using DEGs.

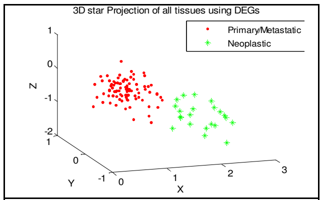


Fig. 6. 3D SCP of Gastric cancer dataset

Case stydy IV: Colon Cancer Dataset

Colon cancer is second most cause of cancerous deaths in western world. Affymetrix oligonucleotide array complementary to more than 6500 human genes are used in this study. The gene expression is studied using 40 tumor samples and 22 normal samples. The preprocessing of this dataset resulted in 2000 interesting genes which have been used as input to Fuzzy-ASI based two-way clustering [2]. The differentially expressed genes with high confidence value are:

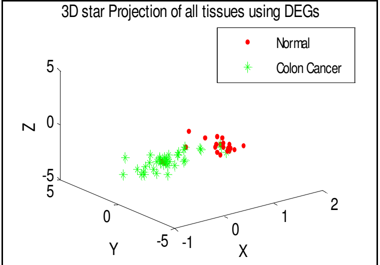


Fig. 7. 3D SCP of Colon cancer dataset

Human cysteine rich protein exons 5 and 6 (100%), human Hsa. 8125 (75%), human desmin gene (100%), human cysteine rich protein (100%), human hmgl mRNA for high mobility group protein Y (90%), uroguanylin precursor (92%), myosin heavy chain nonmuscle (100%), mitochondrial matrix protein precursor (100%), human gene for heterogeneous nuclear ribonucleoprotein (75%), human cysteine rich protein (92%), macrophage migration inhibitory factor (96%), tropomyosin fibroblast and epithelial muscle type (88%), complement factor D precursor (92%), nucleoside diphosphate kinase A (100%), transcription factor IIIA (100%), human serine kinase mRNA complete cds (100%). Myosin regulatory light chain smooth muscle isoform (88%), hevin like protein (96%), H.sapiens mRNA for p cadherin (80%), human vasoactive intestinal peptide(88%), human monocyte derived neutrophil activating protein (76%), human aspartyl tRNA synthetase alpha 2 subunit (96%), collagen alpha 2 (96%) and gelsolin precursor (90%).

Analysis on colon cancer dataset indicates several genes with common function such as 6 ribosomal proteins, 2 muscle related proteins, 2 mitochondrial proteins and several membrane proteins. Other important genes include heat shock protein, human cysteine rich proteins, Desmin genes, integrin associated protein, Human nucleolar protein and complement factor D etc.

Fig. 7 shows the projection of tissues using 3D SCP. As shown in Fig. 7, ALL and AML tissue cases clearly separate using DEGs as features.

Case Study V: Leukemia Dataset

Leukemia is the leading cause of death due to malignant diseases in children in the US. Gene expressions of approximately 6817 genes are used to classify two types of acute Leukemia (ALL and AML). The data consists of 47 (38 B-cell and 9 T-cell) cases of ALL and 25 cases of AML. The data is divided into a training class containing 38 samples and a test class containing 34 samples. The class labels for training class are available from the author [31]. The pre-processing proposed by author resulted in 3571 genes, the rest of the genes are eliminated. The data is further separated into training and test classes. Using the training data, the Fuzzy-ASI algorithm is applied to identify the genes that maximally differentiate between the two classes (ALL and AML). These DEGs are further used as features for clustering the test tissue cases. Empirical analysis shows that DEGs used as features separated the different tissue cases and identified tissues from test case correctly. The differentially expressed genes with high confidence value are:

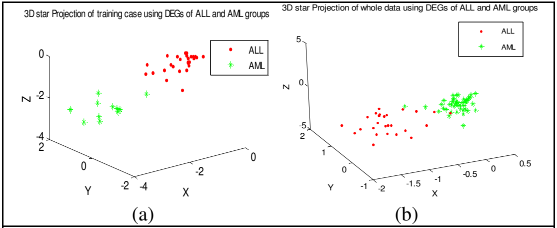


Fig. 8. 3D SCP Visualization of (a) Training class only and (b) Both training and test class.

KIAA0097 gene(100%), KIAA0195 gene(78%), tenascin (78%), cpg-Enriched Dna clone E06 (100%), Glacinamide Ribonucleotide synthetase(100%), bactericidal Bpi Gene (92%), pepsinogen a precursor (75%), HIV1 tata element modulatory factor (85%), induced myelpid leukemia cell differentiation protein MCL1 (100%), ly-9 mRNA, phosphomevalonate kinase mRNA(100%), large praline-rich protein BAT3 (75%), alpha-1 collagen type II gene (95%), GTF2B general transcription factor IIB (78%), actinin alpha 2(75%), insulin activator factor(90%), clone A9A2BR11(CAC)n.(GTG) (78%), peroxisomal enoyl-CoA hydratase like protein (78%), EB1 mRNA(85%), ZNF177 KRAB zinc finger protein (75%), IAP homolog B(MIHB) mRNA (85%), guanine nucleotide binding protein(78%) and D53(95%).

Analysis revealed several genes that may be involved in Leukemia progression. Some of these genes are KIAA0097, HIV1-tata modulator y factor, RBL1, D53, MCL1, Insulin activator factor, malignant melanoma metastasis (KiSS-1), EB1 and putative transmembrane protein etc.

As shown in Fig. 8 (a), different tissue cases clearly separate using DEGs found from training set. The DEGs are further used to categorize the unlabeled tissue test cases. As shown in Fig. 8(b), the DEGs clearly separated different tissue cases for the test case.

IV. CONCLUSTIONS AND FUTURE DIRECTIONS

A two-way fuzzy-ASI based clustering for knowledge discovery from microarray data is developed and empirically evaluated. The goal here is robust selection of DEGs and functionally classifying the genes. It is widely believed that one gene may be involved in more than one cellular function or biological process. To take this into account, the ASI-based crisp clustering algorithm is modified to accommodate soft clustering. This is achieved by introducing the notion of confidence measure on the samples by introducing fuzzy membership. The ASI model with relevance of each gene to each cluster is also explored to answer some important biological questions (cf. section III).

A progressive framework is used for two-way clustering of the microarray data using fuzzy-ASI algorithm. Empirical analysis shows that the fuzzy ASI-based algorithm is robust against arrangement of samples and offers a meaningful clustering result. The co-expressed genes which are also highly differentially expressed are highly ranked. The simulation is repeated 25 times and the number of times a gene is differentially expressed is noted. The high ranked genes found to be highly differentially expressed are further studied by 3D SCP to see if they perform well in separating different tissue cases correctly.

The robustness of fuzzy-ASI based two-way clustering algorithm in identifying the DEGs is validated by using 100 artificial datasets. Further analyses were also performed on a number of real cancer datasets. The analysis of all the three cancer datasets showed some association with tumor generating proteins such as P53 and HIV-tat. It is also seen that most of the genes involved in all the three cancers are associated with metal ion binding, DNA and protein synthesis, cell differentiation, apoptosis, cell proliferation and metabolic activities, a few to mention.

The fuzzy-ASI based two-way clustering algorithm as of now provides the user with functionally classifying genes and differentially expressed genes. It does not explicitly state the relations among the genes in the sense that ‘if the gene X is over expressed, how the gene Y expresses?’ These relations among the genes may be of interest and may provide interesting insight into the functionalities of genes. Such a relational map depicting the dependence-relations among several interesting genes will be explored in future research.

ACKNOWLEDGMENTS

The authors acknowledge the Herff College of Engineering Fellowship program and the faculty start-up grants from The University of Memphis for partially funding this research.

References

[1] I. Guyon, "An Introduction of Variable and Feature Selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157-1182, 2003.

[2] U. Alon, N. Barkai, D. A. Notterman, K.Gish, S. Ybarra, D. Mack, and A. J. Levine, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," *Proc. Natl. Acad. Sci. USA*, vol. 96, pp. 6745-6750, 1999.

[3] K. J. Antonellis, "Optimization of an External Standard for the Normalization of Affymetrix GeneChip Arrays," GeneLogic Inc 2002.

[4] S. Busygin, G. Jacobsen, and E. Kramer, "Double Conjugated Clustering Applied to Leukemia Microarray Data," *2nd SIAM, ICDM, Workshop on clustering high dimensional data*, 2002.

[5] A. Califano, G. Stolovitsky, and Y. Tu, "Analysis of gene expression microarrays for phenotype classification," *Proceedings of international conference on Intelligent Systems in Molecular Biology*, vol. 8, pp. 75-85, 2000.

[6] G. Getz, E. Levine, and E. Domany, "Coupled two-way clustering of gene microarray data," *Proceedings of National Academy of Science, USA*, vol. 97, pp. 12079-12084, 2000.

[7] A. Pascual-Montano, F. Tirado, P. Carmona-Saez, J. M. Carazo, and R. D. Pascual-Marqui, "Two-way clustering of gene expression profiles by sparse matrix factorization," *IEEE computational Systems Bioinformatics Conference Workshops (CSBW 05)*, pp. 103-104, 2005.

[8] C. Tang and A. Zhang, "Interrelated Two-way Clustering: an unsupervised approach for gene expression data analysis," *In Proceedings of the 2nd IEEE international Symposium on Bioinformatics and Bioengineering*, vol. 14, pp. 41-48, 2001.

[9] S. Y. Kim, J. W. Lee, and J. S. Bae, "Effect of Data Normalization on Fuzzy Clustering of DNA microarray Data," *Bioinformatics*, vol. 7, 2006.

[10] R. D'andrade, "U-Statistic Hierarchical Clustering," *Psychometrika*, vol. 4, pp. 58-67, 1978.

[11] S. C. Johnson, "Hierarchical Clustering Schemes," *Psychometrika*, vol. 2, pp. 241-254, 1967.

[12] Y. Zhao and G. Karypis, "Evaluation of hierarchical clustering algorithms for document datasets," Dept. Of Comp Sc., University of Minnesota.

[13] T. Kohonen, "Self-Organizing Maps," *Springer Series in Information Sciences*, vol. 30, 1995.

[14] R. O. Duda, P. E.Hart, and D. G.Stork, *Pattern Classification*, 2nd ed: John Wiley and Sons Inc, 2000.

[15] D. Dembele and P. Kastner, "Fuzzy C-means Method for Clustering Microarray Data," *Bioinformatics*, vol. 19, pp. 973-980, 2003.

[16] W. Yang, L. Rueda, and A. Ngom, "A Simulated Annealing Approach to Find the Optimal Parameters for Fuzzy Clustering Microarray Data," *Chilean Computer Science Society*, 2005.

[17] T. Li, S. Ma, and M.Ogihara, "Document Clustering via Adaptive Subspace Iteration," *Special Information Group on Information Retrieval 2004*, pp. 218-225, 2004.

[18] J. Shaik and M. Yeasin, "Functionally Classifying Genes from Microarray Data Using Linear and Non-linear Data Projection," *The 4th ACS/IEEE International Conference on Computer Systems and Applications*, 2006.

[19] J. Shaik and M. Yeasin, "A Progressive Framework for Two-Way Clustering Using Adaptive Subspace Iteration for Functionally Classifying Genes," *Proceedings of IEEE IJCNN'06, Vancouver, Canada.*, pp. 5287-5292, 2006.

[20] D. L. Davies and D. W. Bouldin, "A Cluster Separation Measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 1, pp. 224-227, 1979.

[21] J. Shaik and M. Yeasin, "Adaptive Ranking and Selection of Differentially Expressed Genes from Microarray Data," *WSEAS transactions on Biology and Biomedicine*, vol. 3, pp. 125-133, 2006.

[22] G. J. McLachlan, R. W. Bean, and D. Peel, "A mixture model-based approach to the clustering of microarray expression data," *Bioinformatics*, vol. 18, pp. 413-422, 2002.

[23] K. S. Pollard and M. J. v. d. Laan, "Statistical Inference for Simultaneous Clustering of Gene Expression Data," *Mathematical BioSciences*, vol. 176, pp. 9121-9126, 2002.

[24] J. Shaik and M. Yeasin, "Visualization of High Dimensional Data using an Automated 3D Star Co-ordinate System," *Proceedings of IEEE IJCNN'06, Vancouver, Canada.*, pp. 2318-2325, 2006.

[25] I. Lonnstedt and T. Speed, "Replicated Microarray Data," *Statistica Sinica*, vol. 12, pp. 31-46, 2002.

[26] S. Mukherjee, S. J. Roberts, and M. J. Laan, "Data-adaptive Test Statistics for Microarray Data," *Bioinformatics*, vol. 21, pp. 108-114, 2005.

[27] J. Shaik and M. Yeasin, "Performance Evaluation of Subspace-based Algorithm in Selecting differentially Expressed Genes and Classification of Tissue Types from Microarray Data," *Proceedings of IEEE IJCNN'06, Vancouver, Canada.*, pp. 5279-5286, 2006.

[28] J. Shaik and M. Yeasin, "A Unified Framework for Knowledge Discovery in Cancer Microarray Datasets," *submitted to IEEE Transactions on Biology and Biomedicine*, 2006.

[29] J. Shaik and M. Yeasin, "Ranking Function Based on Higher Order Statistics (RFHOS) for Microarray Data Analysis," *Submitted to RECOMB*, 2007.

[30] X. Chen, S. Y. Leung, S. T. Yeuen, K. M. Chu, J. Ji, R. Li, A. S. Y. Chan, S. Law, O. G. Troyanskaya, J. Wong, S. So, D. Botstein, and P. O. Brown, "Variation in Gene Expression Patterns in Human Gastric Cancers," *Mol Bio Cell*, vol. 14, pp. 3208-3215, 2003.

[31] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, pp. 531-537, 1999.