Mining Robust Overlapping Co-Clustering in the Presence of Noise

P. Sudhakar, PhD
Assistant Professor,
Dept. of Computer Science and Engineering,
Annamalai University,
Chidambaram, India.

S. Saranya
P.G Student,
Dept. of Computer Science and Engineering,
Annamalai University,
Chidambaram, India

ABSTRACT

Data clustering techniques have been applied to extract information from gene expression data for two decades. A large volume of novel clustering algorithms have been developed and achieved great achievement. However, due to the various structures and intensive noise, there is no reliable clustering approach can be applied to all gene expression data. In this paper, the problem of revealing robust overlapping co-clustering is identified in the presence of noise. Instead of requiring all objects in a cluster have identical attribute order, this system requires that (1) at least a certain fraction of the objects have identical attribute order; (2) other objects in the cluster may deviate from the consensus order by up to a certain fraction of attributes.

Keywords

Rocc, Opc, Aopc Clusters, Gene Expression Data Mining.

1. INTRODUCTION

An array is an orderly arrangement of samples where matching of known and unknown DNA samples is done based on base pairing rules. Thousands of spotted samples known as probes (with known identity) are immobilized on a solid support (a microscope glass slides or silicon chips or nylon membrane). A skin condition can be DNA, cDNA, or oligonucleotides. These are used to determine opposite binding of the unknown sequences thus allowing parallel analysis for gene expression and gene discovery. A research with a single DNA chip can provide information on thousands of genes simultaneously. A neat arrangement of the probes on the support is important as the location of each spot on the array is used for the identification of a gene.

1) Microarrays allow researchers to view transcription levels and the expression of specific genes during various stages as well as various situations.

2) Microarrays are generally twenty-five base pairs in length.

Understanding the expression of particular genes during developmental stages allows a more thorough study of diseases as well as their response to treatment options.

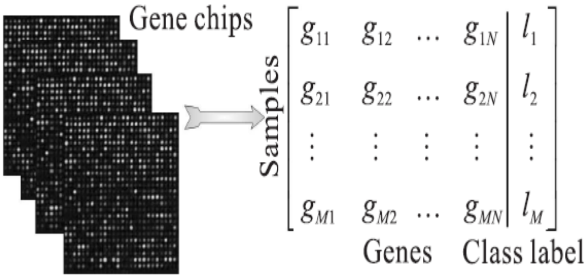

Gene chips
Samples
g_{11} g_{12} … g_{1N} l_1
g_{21} g_{22} … g_{2N} l_2
g_{M1} g_{M2} … g_{MN} l_M
Genes
Class label

Fig. 1: Matrix presentation of a gene expression dataset

Clustering is the process of grouping a set of objects into classes of similar objects. The similar objects are required to have close values on at least a set of dimensions. Clustering has been extensively studied in many areas, including figures, machine learning, pattern recognition, and image processing. Recent efforts in data mining have focused on methods for efficient and effective cluster analysis in large databases. Clustering high dimensional data using traditional algorithms has suffered from the fact that many attributes may be irrelevant and can thus mask clusters located in some subspaces. Subspace clustering algorithms have freshly been proposed to solve this problem. They search for clusters in subspaces formed by relevant attributes. Among various subspace clustering models, one was planned to mine a set of objects which show identical attribute order, called Order Preserving Cluster (OPC). This model originally attracts researchers' interests because of its important utility in gene expression data analysis. Finding such local expression patterns exhibited under relevant conditions is one important contribution of the OPC algorithm and may be the key to uncover significant but previously unknown genetic pathways.

Noise is ubiquitous in real data due to technical errors, missing values and variable experimental situation, etc. The underlying OPCs may be broken into small ones by noise and cannot be captured by any strict model due to their vulnerability to noise. Mining subspace OPCs in the presence of noise is very challenging for the following reasons. First, the search space is often huge due to the curse of dimensionality. For a dataset with n attributes, there are totally candidate subspaces. For OPC mining, the difficulty is much higher, since in addition to identify subspaces, we also need to distinguish different orders.

To our best knowledge, no previous work has been done on the problem of subspace OPC mining with noise tolerance. In this project, we study this problem and suggest a new model. Experimental results show that our new model can capture OPCs contaminated by noise and thus is more robust to noise than the previous severe model. We also propose an algorithm to mine clusters under the new model. Although we deal with a much more challenging problem, our algorithm can find out more significant clusters that cannot be found by the previous strict model in an efficient way.

Order Preserving Cluster is found to be significant in view of the following points:

- Provides a hierarchical cluster solution.
- free from the use of nearness measures.
- faster processing due to basic matching mechanism.
- capable of handling noisy datasets.
- does not require the number of clusters a priori.

2. RELATED WORK

Chun Tang, Li Zhang , Aidong Zhang and Murali Ramanathan [1] they proposed, DNA arrays can be used to measure the expression levels of thousands of genes at the same time. Currently most research focuses on the understanding of the meaning of the data. We present a new construction for invalid analysis of gene expression data which applies an unified joint clustering approach on the gene appearance matrices. The goal of clustering is to find significant gene patterns and execute cluster detection on samples.

Jinze Liu and Wei Wang [2] Clustering is process of alignment a set of object into classes of related items. since of unknowingness of the secreted pattern in the data sets. pending lately, comparison procedures are naturally based on distances, e.g. Euclidean distance and cosine distance. In this scheme, we suggest a flexible yet authoritative clustering model, namely OP-Cluster (Order Preserving Cluster). Under this new model , two objects are related on a separation of size if the rate of these two objects consist of the same relative order of those dimensions. Such a cluster may occur when the appearance levels of (co-regulated) genes can rise or fall synchronously in reaction to a series of location stimulus. Therefore, find of OP-Cluster is necessary in instructive considerable genetic substance inflexible networks.

Amir Ben- Dor, Benny Chor, Richrd Karp and Zohar Yakhini [3] they have proposed, scheme concerns the detection of patterns in gene appearance matrices, in which every part gives the expression stage of a known gene in a particular test. Mainly existing methods for model finding in such matrices be based going on clustering gene with comparing their expression levels in all experiment, or clustering experiment by comparing their expression levels for every genes. Our effort goes further than such overall approaches by looking for limited patterns to apparent them after us focal point at the same time on a subset G of the genes and a subset T of the experiments. specially, we look for order preserving submatrics (OPSMs), in which the expression levels of all genes bring the identical linear ordering of the experiments.

Lance Parsons, Ehtesham Hague and Huan Liu [4] Subspace clustering is an expansion of usual clustering that seeks to discover clusters in similar subspaces inside a dataset. Element choice removes unrelated and unneeded size by analyzing the whole dataset. Subspace clustering algorithms restrict the search for appropriate magnitude allowing them to discover clusters that be in several, maybe overlapping two main undergrowth of subspace clustering based on their explore approach. Top-down algorithms discover a first clustering in the complete set of dimensions and calculate the subspaces of each cluster, iteratively civilizing the results.

Hualong Yu,Jun Ni,Yuanyuan,Dan and Sen Xu [5] present have been many distorted cancer gene expression datasets in the post-genomic era. removal of disparity expression genes or creation of resolution system using these distorted datasets by traditional algorithms will acutely undervalue the presentation of the alternative class, most important to incorrect analysis in medical trails. This paper presents a skewed gene assortment algorithm that introduces a slanted metric into the gene assortment process. The extracted genes are matching as choice rules to differentiate both classes, through these judgment rules then included into an assembly knowledge structure by popular determination to identify test examples; thus avoiding monotonous data normalization and classifier production. The withdrawal and integrate of a few consistent result rules gave superior or at least similar arrangement presentation than several traditional class discrepancy training algorithms on four standard excessive cancer gene appearance datasets.

3. OUTLINE OF THE WORK

The rest of the paper is prepared as follows. In Section 2, we review some related work. Our AOPC model is presented in Section 3. Section 4 explains proposed ROCC clustering algorithm in detail. Experimental results are shown in Section 5 and we conclude the paper in Section 6.

4. PROPOSED APPROACH

Robust Overlapping Co-clustering (ROCC) is based on a systematically developed objective function, which is minimized by an iterative method that provably converges to a in the neighborhood best solution. ROCC is also robust to the noise model of the data and can be modified to use the most right distance calculate for the data, selected from a big class of distance calculate.
This approach is robust in the presence of noisy and automatically objects as well as features. This algorithm reluctantly detects and prunes during the clustering process.

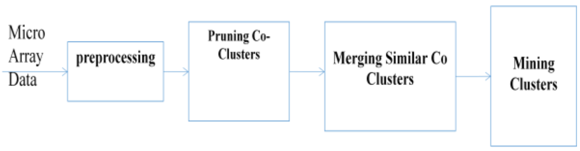


Fig. 2: Block diagram of proposed system

5. MODULES

In the proposed system is divided into 4 modules.

Module 1: Preprocessing

Module 2: Pruning Co-Clusters

Module 3: Merging Co-Clusters

Module 4: Mining Co-Cluster

These modules are explained as follows.

5.1 Preprocessing

Getting Gene Matrix from the Comma Separated Dataset is the first stage of the project. Removing the empty or null values from the dataset is the important task to find the systematic clusters. In the gene expression matrix, different genes have different ranges of intensity values. The intensity values only cannot include significant meaning, Other than the relative values are more inherent. So we first normalize the original gene intensity values into relative values. Our common method is,

$$w'_{i,j} = \frac{w_{i,j} - \mu_i}{\mu_i} \text{ where, } \mu_i = \frac{\sum_{j=1}^{m} w_{i,j}}{m}$$

W'$_{i,j}$ denotes normalized intensity value for gene i of sample j, W$_{i,j}$ represents the original intensity value for gene i of sample j, m is the number of samples, and u$_i$ is the mean of the intensity values for gene i over all samples. Note that along with thousands of genes, not every one of them has the equal part in individual the classes. In fact, some genes have small part. We must to remove individual genes which have small reaction to the experiment condition. We consider genes whose intensity values go on invariant or alter very small belong to this class refer to each gene vector (after normalization) as, Gi = ($w_{i,1}$,$w_{i,2}$,$w_{i,3}$,… $w_{i,n}$), Where i=1,2,…n for each gene.

5.2 Pruning Co-Clusters

Filtering and reduces the calculation crack necessary by the following merging step. But one has no idea of the final number of co-clusters,

A simple and efficient filtering heuristic is to select the error cut-off value as the one at which the sorted co-cluster errors show the largest increase among repeated values. The co-clusters with errors greater than the cut-off are filtered out.

5.3 Merging Co-Clusters

Every agglomeration identifies the "closest" pair of co-clusters that can be well represented by a single co-cluster model and are thus possibly part of the equal original co-cluster, and merges them to form a new co-cluster. The rows and columns of the new co-cluster consist of the joining together of the rows and columns of the two merged co-clusters.

Merging co-clusters in this method allows co-clusters to distribute rows and columns and thus allows incomplete overlap connecting co-clusters. The increase in the distance between successively merged co-clusters is then computed and the set of co-clusters just before the largest increase is selected as the final solution.

5.4 Mining CO-Cluster

Here every heterogeneous group, first choose one sample, and then use the remaining samples of this group to select important genes, and calculate the class of the pending samples.

The process is repeated for every sample, and the increasing error rate is calculated. Once the heterogeneous group which has lower error rate is found, its matching reduced gene sequence is selected as G with n2 genes for the next iteration.

After previous step, the gene number is reduced from n1 to n2. The above steps can be repeated by clustering n2 genes, and so on. The iteration will be terminated pending the termination conditions are satisfied.

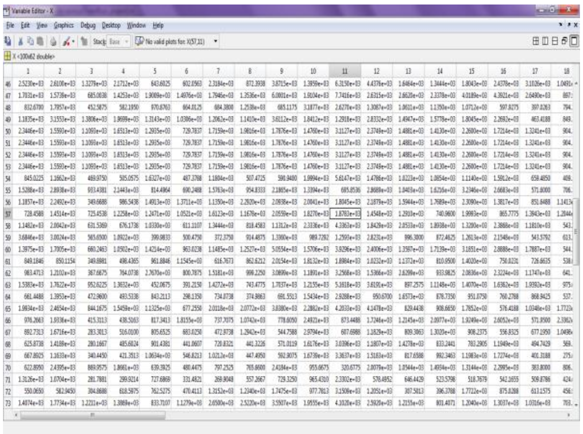
6. EXPERIMENTAL RESULTS

In this part, we study the performance of our algorithm through a sequence of experiments. To make the experiment results more sensible and thus believable, we conduct all experiments on a real gene expression dataset. This dataset is the yeast cell cycle data]. Each row of this dataset records the expression levels across 18 time points for a gene.

Totally 799 genes are in this dataset. The input parameters to the mining results and provide some guidance on how to choose them properly. There are totally four parameters that need to be specified in our algorithm. smin and lmin are set to define the valid AOPC. In Phase 2, δc, δs are set to define the ROCC. In principle, the optimal values of these parameters depend on the size and shape of the salient clusters and the level and distribution of noise in the dataset. For our dataset, we use smin=60, lmin=5 and found 682 AOPCs in 495 seconds. As for δc and δs, they are often set to some moderate level to control the noise tolerance. As a rule of thumb, δs should be no larger than 0.5, otherwise it is too strict to prevent ROCCs from being merged; δc should be in [0.6, 0.8], since a too small value tends to include too much noise and thus decreases the significance of the results.

6.1 Result

In Fig. 3 Preprocessed Dataset in Fig. 4 Preprocessed Dataset Plotting in Fig. 5 Relative Expression Level for Each Gene Dataset in. Fig. 6 First Level of Cluster in Fig.7 Pruning Co-Cluster Diagram in Fig. 8 Mining Co-Cluster Diagram Finally Fig. 9 Cluster Creation.



X <100x62 double>

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
46	2.5230e+03	2.6100e+03	1.3279e+03	2.1712e+03	643.6025	602.0563	2.3184e+03	872.3938	3.8715e+03	1.3959e+03	6.3150e+03	4.4376e+03	1.6464e+03	1.3444e+03	1.8043e+03	2.4378e+03	3.1026e+03	1.0491
47	1.7031e+03	1.5736e+03	685.0038	1.4253e+03	1.9009e+03	1.4976e+03	1.7946e+03	1.3536e+03	6.0001e+03	1.9104e+03	7.7416e+03	2.6315e+03	2.6620e+03	2.3378e+03	4.0189e+03	4.3921e+03	2.6490e+03	897.
48	832.6700	1.7957e+03	452.5875	582.1950	970.8763	664.0125	684.3800	1.2538e+03	685.1175	3.1877e+03	2.6270e+03	1.3087e+03	1.0611e+03	1.1350e+03	1.0712e+03	597.8275	397.0263	794.
49	1.1835e+03	3.1553e+03	1.3806e+03	1.9699e+03	1.3143e+03	1.0306e+03	1.2062e+03	1.1410e+03	3.6112e+03	1.8412e+03	1.2918e+03	2.8332e+03	1.4947e+03	1.5778e+03	1.8045e+03	2.2692e+03	463.4188	849.
50	2.3446e+03	1.5593e+03	1.1093e+03	1.6513e+03	1.2935e+03	729.7837	1.7159e+03	1.9816e+03	1.7876e+03	1.4760e+03	3.1127e+03	2.3749e+03	1.4881e+03	1.4130e+03	2.2600e+03	1.7214e+03	1.3241e+03	904.
51	2.3446e+03	1.5593e+03	1.1093e+03	1.6513e+03	1.2935e+03	729.7837	1.7159e+03	1.9816e+03	1.7876e+03	1.4760e+03	3.1127e+03	2.3749e+03	1.4881e+03	1.4130e+03	2.2600e+03	1.7214e+03	1.3241e+03	904.
52	2.3446e+03	1.5593e+03	1.1093e+03	1.6513e+03	1.2935e+03	729.7837	1.7159e+03	1.9816e+03	1.7876e+03	1.4760e+03	3.1127e+03	2.3749e+03	1.4881e+03	1.4130e+03	2.2600e+03	1.7214e+03	1.3241e+03	904.
53	2.3446e+03	1.5593e+03	1.1093e+03	1.6513e+03	1.2935e+03	729.7837	1.7159e+03	1.9816e+03	1.7876e+03	1.4760e+03	3.1127e+03	2.3749e+03	1.4881e+03	1.4130e+03	2.2600e+03	1.7214e+03	1.3241e+03	904.
54	845.0225	1.1662e+03	469.9750	505.0675	1.6327e+03	487.3788	1.1804e+03	507.4725	590.9400	1.9994e+03	5.6147e+03	1.4786e+03	1.0223e+03	1.0854e+03	1.1140e+03	1.5912e+03	658.4850	409.
55	1.5288e+03	2.8938e+03	933.4381	2.1443e+03	814.4864	690.2488	1.5763e+03	954.8333	2.1865e+03	1.3394e+03	685.8536	2.8689e+03	1.0403e+03	1.6216e+03	3.2346e+03	2.6683e+03	571.8000	706.
56	1.1857e+03	2.2492e+03	349.6688	986.5438	1.4913e+03	1.3711e+03	1.1350e+03	2.2920e+03	2.0938e+03	2.0041e+03	1.8045e+03	2.1879e+03	1.5944e+03	1.7689e+03	2.3090e+03	1.3817e+03	851.6488	1.1413
57	728.4588	1.4514e+03	725.4538	1.2258e+03	1.2471e+03	1.0521e+03	1.6123e+03	1.1678e+03	2.0559e+03	1.8270e+03	1.8783e+03	1.4548e+03	1.2910e+03	740.9600	1.9993e+03	865.7775	1.3943e+03	1.2844
58	1.1482e+03	2.0042e+03	631.5369	676.1738	1.0330e+03	611.1107	1.3444e+03	818.4583	1.1312e+03	2.3336e+03	4.3363e+03	1.8429e+03	2.8533e+03	1.8938e+03	1.3280e+03	2.3888e+03	1.1810e+03	543.
59	3.6846e+03	3.0024e+03	585.6500	1.0922e+03	399.9833	500.4750	372.3750	914.4875	1.3360e+03	989.7292	1.2593e+03	2.8231e+03	996.3000	872.4625	1.2613e+03	2.1548e+03	543.5792	613.
60	1.3975e+03	1.7005e+03	660.2463	1.0502e+03	1.4214e+03	963.0238	1.1485e+03	1.2527e+03	5.0554e+03	1.5706e+03	3.9296e+03	2.4006e+03	1.3587e+03	1.7139e+03	3.1851e+03	2.8888e+03	1.7887e+03	544.
61	849.1846	850.1154	349.8981	498.4365	961.8846	1.1545e+03	616.7673	862.6212	2.0154e+03	1.8132e+03	1.8984e+03	1.0232e+03	1.1372e+03	810.9500	1.4020e+03	750.0231	726.6635	538.
62	983.4713	1.2102e+03	387.6875	764.0738	2.7670e+03	800.7875	1.5181e+03	999.2250	3.0899e+03	1.1891e+03	3.2568e+03	1.5366e+03	2.6299e+03	933.9825	2.0836e+03	2.3224e+03	1.1747e+03	641.
63	1.5383e+03	1.7622e+03	952.6225	1.3632e+03	452.0675	391.2150	1.4272e+03	743.4775	1.7037e+03	1.2155e+03	5.3618e+03	3.6191e+03	897.2575	1.1148e+03	1.4070e+03	1.6362e+03	1.9392e+03	975.
64	661.4488	1.3953e+03	472.9600	493.5338	843.2113	298.1350	734.0738	374.9863	691.5513	1.5434e+03	2.9288e+03	950.6700	1.6573e+03	878.7350	951.0750	760.2788	868.9425	537.
65	1.9934e+03	2.4654e+03	844.1675	1.5459e+03	1.1325e+03	677.2558	2.0118e+03	2.0772e+03	3.8380e+03	2.2882e+03	4.2033e+03	4.1478e+03	829.4438	908.6650	1.7852e+03	576.4188	1.0348e+03	1.7723
66	976.2663	1.9336e+03	415.3113	438.5063	817.3413	1.8155e+03	737.7075	1.0742e+03	778.6050	2.4921e+03	873.4488	1.7246e+03	1.2145e+03	2.0977e+03	1.9249e+03	2.6052e+03	571.0500	2.3362
67	892.7313	1.6716e+03	283.3013	516.0100	805.6525	683.0250	472.9738	1.2942e+03	544.7588	2.9794e+03	607.6988	1.1829e+03	809.3863	1.3020e+03	908.2375	556.8325	677.1950	1.0498
68	625.8738	1.4189e+03	280.1667	485.6024	901.4381	441.0607	720.8321	441.3226	571.0119	1.6176e+03	3.0396e+03	1.1807e+03	1.4278e+03	833.2441	783.2905	1.1949e+03	494.7429	569.
69	667.8925	1.1633e+03	340.4450	421.3513	1.0634e+03	546.8213	1.0212e+03	447.4950	592.9075	1.6739e+03	3.3637e+03	1.5183e+03	817.6588	992.3463	1.1983e+03	1.7274e+03	401.3188	275.
70	622.8950	2.4395e+03	889.9575	1.8661e+03	639.3925	480.4475	797.2525	765.8600	2.4184e+03	955.6675	320.6775	2.0079e+03	1.8544e+03	1.4934e+03	1.3144e+03	2.2995e+03	383.8000	806.
71	1.3126e+03	1.0704e+03	281.7881	299.9214	727.6869	331.4821	269.9048	557.2667	729.3250	965.4310	2.3302e+03	578.4952	646.4429	523.5798	518.7679	542.1655	509.8786	424.
72	550.0650	582.9450	304.8688	618.5875	762.5275	470.4113	1.3152e+03	1.2340e+03	1.7475e+03	977.7813	3.1508e+03	1.2051e+03	307.5813	396.3788	1.7722e+03	875.0288	613.1575	456.
73	1.4074e+03	1.7734e+03	1.2211e+03	1.3869e+03	833.7107	1.1279e+03	2.6500e+03	2.5220e+03	3.5507e+03	1.9555e+03	4.3020e+03	2.5920e+03	1.2155e+03	801.4071	1.2040e+03	1.3037e+03	1.0316e+03	703.

Fig. 3 Preprocessed Dataset

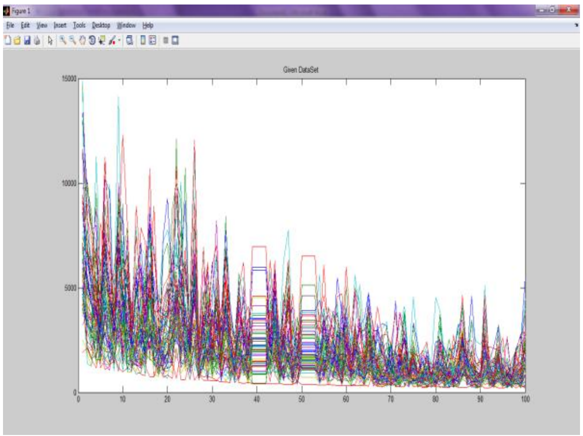


Fig. 4 Preprocessed Dataset Plotting

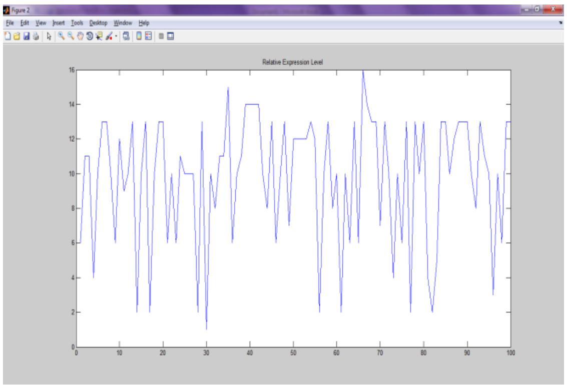


Fig. 5 Relative Expression Level for Each Gene Dataset

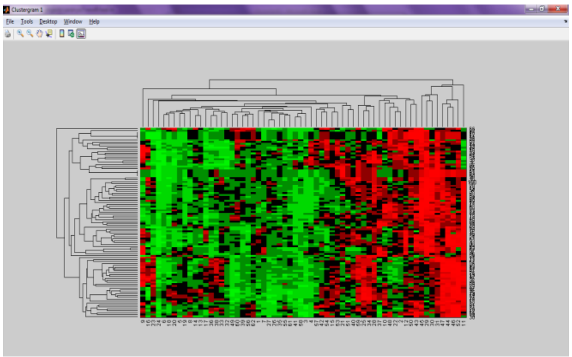


Fig. 6 First Level of Cluster

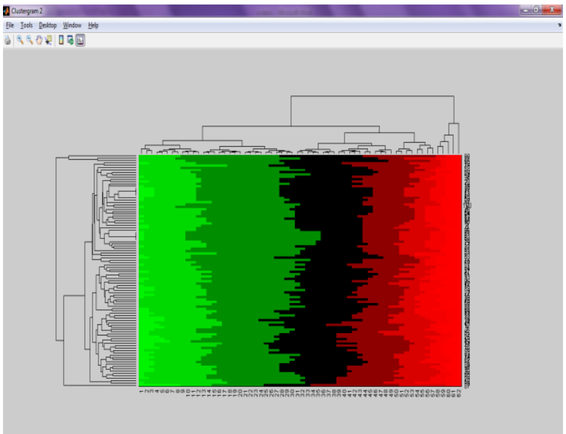


Fig.7 Pruning Co-Cluster Diagram

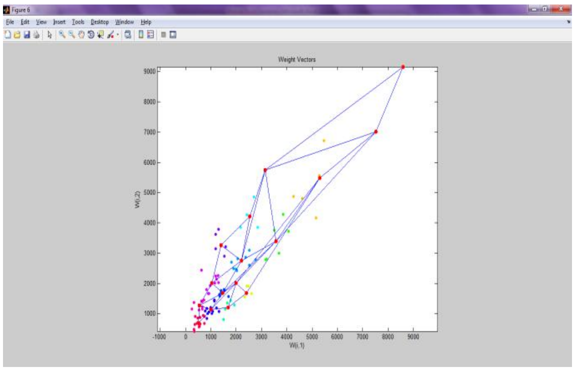
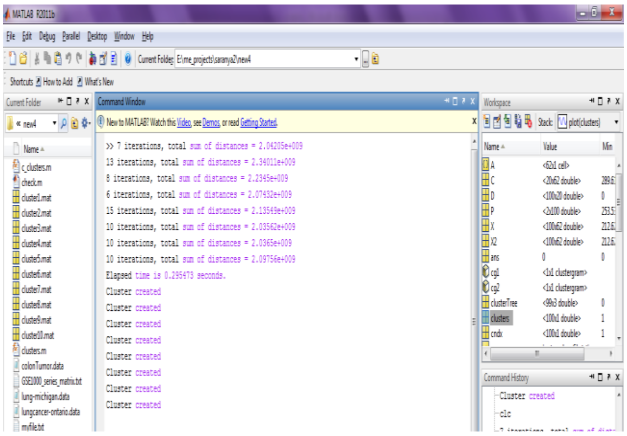


Fig. 8 Mining Co-Cluster Diagram`



New to MATLAB? Watch this Video, see Demos, or read Getting Started.

```
>> 7 iterations, total sum of distances = 2.04205e+009
13 iterations, total sum of distances = 2.34011e+009
8 iterations, total sum of distances = 2.2345e+009
6 iterations, total sum of distances = 2.07432e+009
15 iterations, total sum of distances = 2.13549e+009
10 iterations, total sum of distances = 2.03562e+009
10 iterations, total sum of distances = 2.0365e+009
10 iterations, total sum of distances = 2.09756e+009
Elapsed time is 0.295473 seconds.
Cluster created
Cluster created
Cluster created
Cluster created
Cluster created
Cluster created
Cluster created
Cluster created
```

Fig. 9 Cluster Creation

7. CONCLUSION AND FUTURE ENHANCEMENT

A new model called Robust Overlapping Co-Clustering (ROCC) is proven to be more robust to noise than AOPC model. In addition to its robustness, this model also has several good properties which allow us to mine the ROCCs using an efficient greedy algorithm. Proposed pre-filtering technique can be quickly excluding most ROCC pairs that cannot be merged. And also a hierarchical merging scheme is proposed to further improve the completing time. Experiments on real gene expression data demonstrate that these techniques are efficient to speed up the mining process and the ROCCs are more biologically significant than the AOPCs. Note that, even though our experiments are run on gene expression data, the method presented in this system can be used widely in many applications beyond gene expression analysis.

Main objective of this future work is use to Robust Overlapping Co-Clustering (ROCC). A more flexible model which can tolerate noise is needed.

8. REFERENCES

[1] Chun Tang, Li Zhang, Aidong Zhang and Murali Ramanathan. Interrelated Two-way Clustering: An Unsupervised Approach for Gene Expression Data Analysis 2011.

[2] Jinze Liu and Wei Wang. OP-Cluster: Clustering by Tendency in High dimensional Space 2004.

[3] Amir Ben- Dor, Benny Chor, Richrd Karp and Zohar Yakhini .Discovering Local Structure In Gene Expression Data: The Order – Preserving Sub matrix Problem 2003.

[4] Haixun Wang, Wei Wang, Jiong Yang, Philip S. Yu. Clustering by Pattern Similarity in Large Data Sets 2007. Arabinda Panda, Satchidananda Dehuri. Biclustering for Microarray Data: A Short and Comprehensive Tutorial 2012.

[5] Lance Parsons, Ehtesham Hague and Huan Liu. Subspace Clustering for High Dimensional Data 2004.

[6] J.Z. Liu and W. Wang, "OP-Cluster: Clustering by tendency in high dimensional space", In *Proceedings of the 3rd IEEE International Conference on Data Mining (ICDM)*, 2003, pp. 187-194.

[7] L. Bergroth, H. Hakonen and T. Raita, "A survey of longest common subsequence algorithms", SPIRE, pp 39-48, 2000.

[8] Spellman et al., "Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization", Molecular Biology of the Cell 9, 1998, pp. 3273-3297.

[9] Fisher, R.A.(1922), "On the interpretation of $\chi 2$ from contingency tables, and the calculation of P", *Journal of the Royal Statistical Society* 85(1): pp. 87-94.

[10] L. Parsons, E. Haque, and H. Liu, "Subspace clustering for high dimensional data: a review", ACM SIGKDD Explorations Newsletter, Volume 6, Issue 1, 2004, pp. 90-105.

[11] Ben-Dor, B. Chor, R.M. Karp and Z. Yakhini, "Discovering local structure in gene expression data: the order-preserving submatrix problem", Journal of Computational Biology 10(3/4), 2003, pp. 373-384.

[12] Y. Cheng and G.Church, "Biclustering of expression data", In Proceedings of the 8th International Conference on Intelligent System for Molecular Biology, 2000.

[13] H. Wang, W. Wang, J. Yang and P. Yu, "Clustering by pattern similarity in Gene ontology consortium, www.geneontology.org.

9. AUTHORS PROFILE

Dr. P. Sudhakar obtained him Bachelor's degree in Electronic Communication & Engineering from Bharathidasan University, Tamilnadu, India in 2000. Then he obtained him Master's degree in Computer Science & Engineering from Annamalai University, Tamilnadu, India in 2005. He obtained him ph.d in Computer Science & Engineering from Anna University, Tamilnadu, India in 2012. He is currently working as an Assistant Professor in the Department of Computer Science & Engineering, Faculty of Engineering & Technology, Annamalai University. He is having 15 years of experience in teaching.

S.SARANYA obtained her Bachelor's degree in Computer Science from Annamalai University, Tamilnadu, India in 2013. Now she is doing her Master's degree in Computer Science & Engineering from Annamalai University, Tamilnadu, India.