ANALYSING GENE EXPRESSION USING DATA MINING TECHNIQUES

Dr..CHANDRABOSE1, T. MANIVANNAN2, M. JAYAKANDAN3, P. ANANTHI4
ASSOCIATE PROFESSOR,ASSISTANT PROFESSOR ASSISTANT PROFESSOR ASSISTANT PROFESSOR
EDAYATHANGUDY G.S.PILLAY ARTS & SCIENCE COLLEGE-NAGAPATTINAM-611 002 .

Abstract: With the rapid development of microarray chip technology, gene expression data are being generated in large throughput .The indispensable task of data mining, as a result, is to effectively and efficiently extract useful biological information discussed above from gene expression data. However, the high-dimensionality and the complex relationships among genes impose great challenges for existing data mining methods. Extensive experimental studies are conducted on synthetic and real-life datasets. The experimental results show the effectiveness and efficiency of our algorithms. While we mainly use gene expression data in our study, our algorithms can also be applied to high-dimensional data of other domains.

Keywords: Curler and Reg Miner algorithms full space clustering algorithms2 Top-down Clustering micro cluster

I INTRODUCTION

Gene expression is the process of transcribing a gene's DNA sequence into mRNA sequences, which are later translated into amino acid sequences of proteins. The number of copies of produced RNA is called the expression level of the gene. The regulation of gene expression level is considered important for proper cell function. As an effective technology to study gene expression regulation, microarray gene expression profiling uses arrays with immobilized c DNA or oligonu cleotide sequences to measure the quantity of mRNA based on hybridization. Microarray technologies provide the opportunity to measure the expression levels of tens of thousands of genes in cells simultaneously which are correlated with the corresponding protein made either under different conditions or during different time spots. Gene expression profiles generated by microarrays can help us understand the cellular mechanism of biological process. For instance, it provides information about the cancerous mutation of cells: which genes are most responsible for the mutation, how they are regulated, and how experimental conditions can affect cellular function. With these advantages, microarray technology has been widely used in post genome cancer research studies. With the rapid advance of microarray technology, gene expression data are being generated in large throughput so that an imposing data mining task is to effectively and efficiently extract useful biological information discussed above from the huge and fast-growing gene expression data.

II NONLINEAR CORRELATION AND SHIFTING-AND-SCALING CORRELATION

For high-dimensional data like gene expression data, a subset of data objects (genes) is probably strongly correlated only in a subset of conditions, while not correlated at all in the remaining ones. Besides, the orientation of these local correlation clusters can be arbitrarily oriented. The above problems have been addressed by several subspace clustering algorithms such as LDR, ORCLUS, and 4C are proposed to identify local correlation clusters with arbitrary orientations, assuming each cluster has its own fixed orientation.

Both the linear correlation and the nonlinear correlation subspace clustering methods are density-based, requiring gene members to be close to each other in correlated subspace. However, correlated genes don't need to be close in correlated subspaces at all: positive-correlated genes and negative-correlated genes exhibit no spatial proximity; genes co-regulated together may exhibit pure shifting or pure scaling patterns across the subset of the correlated samples, as addressed in pCluster and TRICLUSTER .

III CONTRIBUTIONS

Propose the concept of Top KRGs to handle the problems of inefficiency and huge rule number in class association rule mining; to address the problem of rule selection in associative classification, present classifier RCBT based on Top KRGs; design two algorithms, CURLER and Reg-Cluster, for finding nonlinear correlation clusters and shifting-and-scaling correlation clusters in subspace respectively. in particular, make the following contributions.

3.1 Top KRGs

propose the concept of top-k covering rule groups (Top KRGs) for each row of a gene expression dataset and have designed a row-wise mining algorithm to discover the top-k covering rule groups for each row. In this way, numerous rules have been clustered into a limited number of rule groups, bounded by $k * n$, where n is the number of rows of gene expression dataset and k is the user specified parameter. Our algorithm is specially efficient for gene expression data with extremely large number of genes but relatively small number of samples. Extensive experiments on real-life gene expression datasets show that our algorithm can be several order of magnitudes better than FARMER , CLOSET+ and CHARM which uses different sets.

3.2 RCBT

Top KRGs also facilitates rule selection for associative classification. Based on that, combine the *nl* rules generated by the most significant genes from each discovered TopKRGs and further develop a new associative classifier called RCBT. Essentially, our RCBT classifier works in a committee-like way. Each test data is first classified by the main classifier built on rules of the top one covering rule groups for each class; if unclassified, the test data is further passed on to the subsequent ordered classifiers built on the rules from the top 2, 3, ..., j covering rule groups until it is classified or $j == k$.

3.3 CURLER

Detecting nonlinear correlation clusters is quite challenging. Unlike the detection of linear correlation in which clusters are of unique orientations, finding nonlinear correlation clusters of varying orientations requires merging clusters of possibly very different orientations. Combined with the fact that spatial proximity must be judged based on a subset of features that are not originally known, deciding which clusters to be merged during the clustering process becomes a challenge. To avoid the problems discussed above, propose a novel concept called *co-sharing level* which captures both spatial proximity and cluster orientation when judging similarity between clusters.

3.4 Reg-Cluster

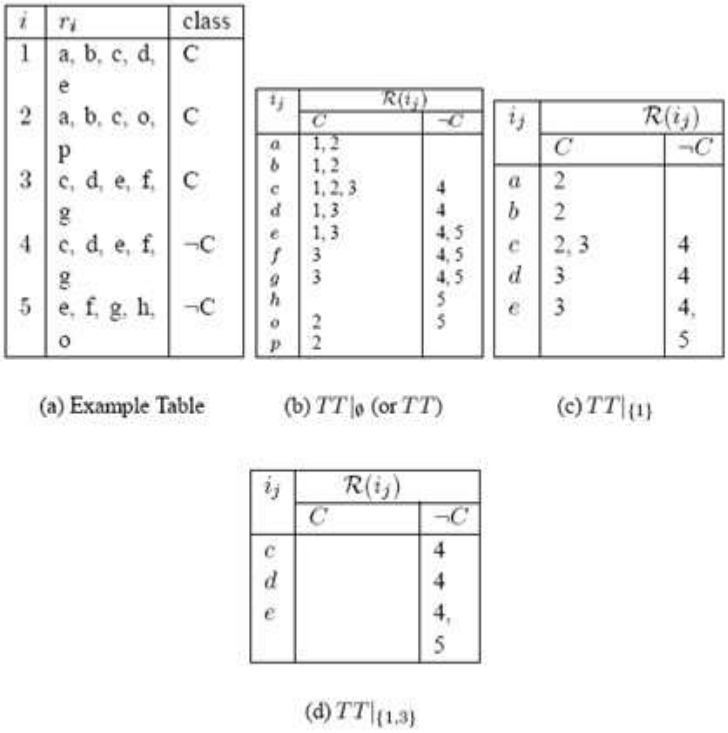
Propose a new model for coherent clustering of gene expression data called **reg-cluster**. The proposed model allows (1) the expression profiles of genes in a cluster to follow any shifting-and-scaling patterns in a certain subspace, where the scaling can be either positive or negative, and (2) the expression value changes across any two conditions of the cluster to be significant, when measured by a user-specified regulation threshold. Also develop a novel pattern-based bi clustering algorithm for identifying shifting-and-scaling co-regulation patterns, satisfying both regulation constraint and coherence constraint.

IV TOPKRGS: EFFICIENT MINING OF TOP K COVERING RULE GROUPS

Define a class association rule as a set of items, or specifically a set of conjunctive gene expression level intervals (*antecedent*) with a single class la- bel (*consequent*). The *general* form of a class association *rule* is: $gene_1[a_1, b_1]$, ...,$gene_n[a_n, b_n] \rightarrow class$, where $gene_i$ is the name of the gene and $[a_i, b_i]$ is its expression interval.

4.1 Problem Statement and Preliminary

To address the problems discussed in the above section, propose to discover the most significant top-*k* covering rule groups (Top kRGS) for each row of a gene expression dataset, will illustrate this with an example.



i	r_i	class
1	a, b, c, d, e	C
2	a, b, c, o, p	C
3	c, d, e, f, g	C
4	c, d, e, f, g	¬C
5	e, f, g, h, o	¬C

(a) Example Table

i_j	$\mathcal{R}(i_j)$	
	C	$\neg C$
a	1, 2	
b	1, 2	
c	1, 2, 3	4
d	1, 3	4
e	1, 3	4, 5
f	3	4, 5
g	3	4, 5
h		5
o	2	5
p	2	

(b) $TT|_{\emptyset}$ (or TT)

i_j	$\mathcal{R}(i_j)$	
	C	$\neg C$
a	2	
b	2	
c	2, 3	4
d	3	4
e	3	4, 5

(c) $TT|_{\{1\}}$

i_j	$\mathcal{R}(i_j)$	
	C	$\neg C$
c		4
d		4
e		4, 5

(d) $TT|_{\{1,3\}}$

Figure 4.1: Running Example

covering rule group for rows r_1 and r_2 is {abc → C} with confidence 100%, the top-1 covering rule group for row r_3 is {cde → C} with confidence 66.7%, and the top-1 covering rule group for rows r_4 and r_5 is {fge → ¬C} with confidence 66.7%. The support values of the above top-1 covering rule groups are all 2, which is equal to minsup.

4.2 Efficient Discovery of Top kRGS

The first problem that address is to efficiently discover the set of top-k covering rule groups for each row (Top kRGS) of gene expression data given a user-specified

4.3 **Experimental Studies**

Dataset	# Genes	# Genes after Discretization	Class 1	Class 0	# Training	# Test
ALL/AML (ALL)	7129	866	ALL	AML	38 (27 : 11)	34
Lung Cancer (LC)	12533	2173	MPM	ADCA	32 (16 : 16)	149
Ovarian Cancer (OC)	15154	5769	tumor	normal	210 (133 : 77)	43
Prostate Cancer (PC)	12600	1554	tumor	normal	102 (52 : 50)	34

Table 4.1: Gene Expression Datasets

4 popular gene expression datasets for experimental studies are used. The 4 datasets are the clinical data on ALL-AML leukemia (ALL) , lung cancer (LC) , ovarian cancer(OC) , and prostate cancer (PC) . In such datasets, the rows represent clinical samples while the columns represent the activity levels of genes/proteins in the samples. There are two categories of samples in these datasets.

4.4 Comparisons of Runtime on Gene Expression Dataset

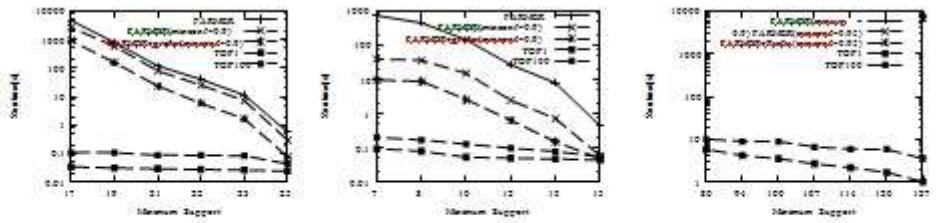


Figure 4.2 (a) ALL-AML leukemia (b) Lung Cancer (c) Ovarian Cancer

V CBT: CLASSIFICATION WITH TOP K COVERING RULE GROUPS

Recent studies have shown that class association rules are very useful in classification. Due to their relative simplicity, they can be easily interpreted by biologists, providing great help in the search for gene predictors (especially those still unknown to biologists) of the data categories (classes). Moreover, it is shown in that classifiers built from association rules are rather accurate in identifying cancerous cell. RCBT is one novel associative classifier built on class association rules.

5.1 RCBT and CBA Classifier

First prove that the set of top-1 covering rule groups for each row contain the set of rules required to build CBA classifier. The basic idea of CBA can be summarized as the following steps:

Step 1: Generate the complete set of class association rules CR for each class that satisfy the user-specified minimum support and minimum confidence.

Step 2: Select rules from sorted rule set CR. For each rule r in CR, if it can correctly classify some training data in D, CBA puts it into classifier C^0, removes those training data covered by r and continues to test the rules after r in CR. Meanwhile, CBA selects the majority class in the remaining data as default class and computes the errors made by current C^0 and default class. This process continues until there are no rules or no training data left.

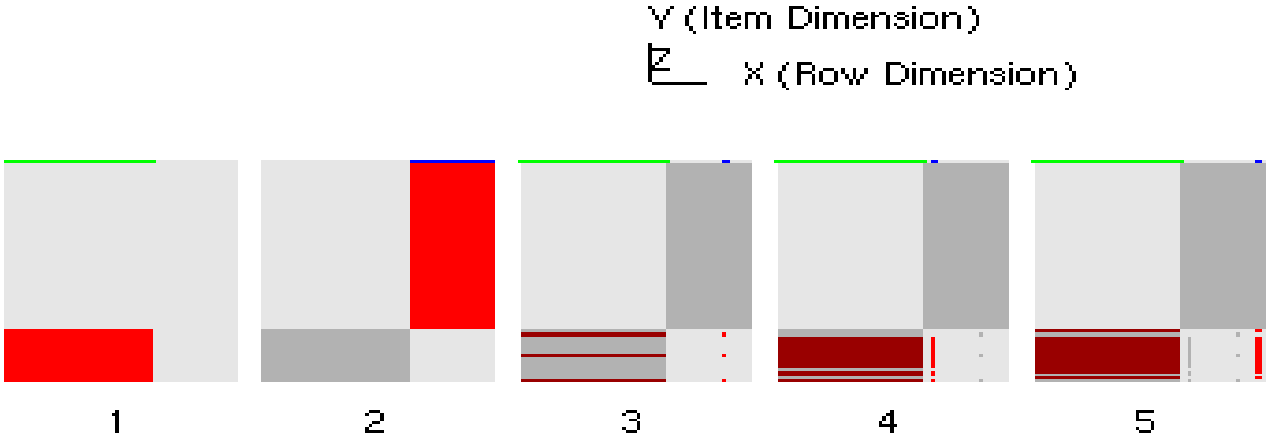


Figure 5.1 RCBT and CBA Classifier

Each node in the lattice except the root node maps to the *antecedent support set* of one rule group in the rule group subset. The *antecedent support set* of the parent node includes that of the child node. The root node corresponds to the set of all the 47 rows.

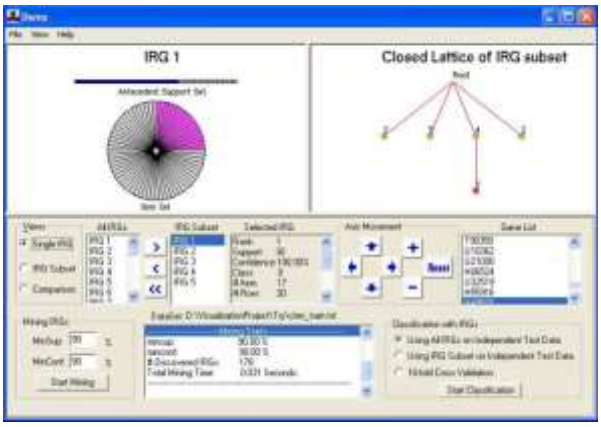
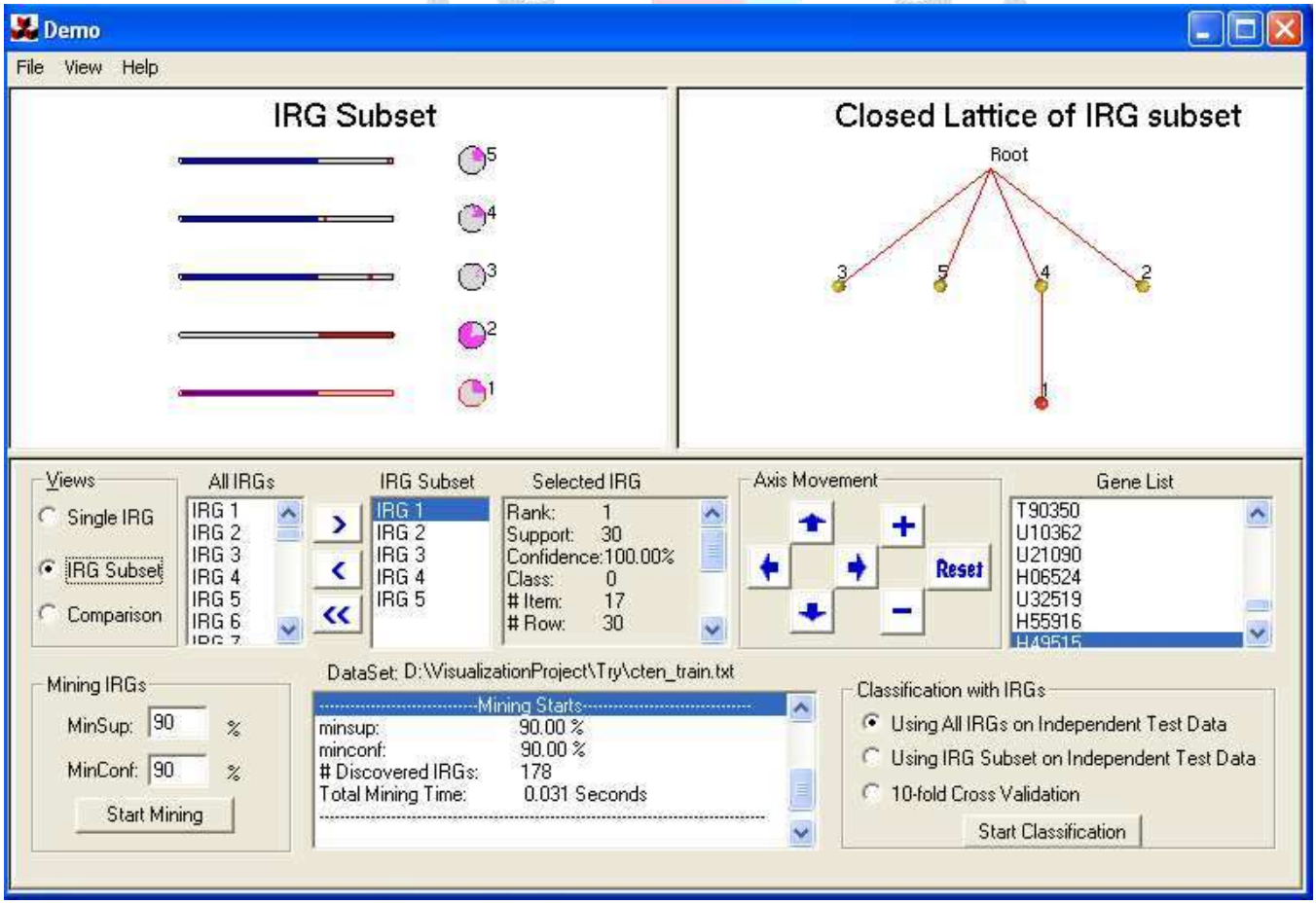


Figure : 5.2 Semantic Visualization of a Single rule group Using the Barcode View and the Flower View

Figure :5.3 Rule Group Comparisons Using the Matrix View

5.2 RCBT Classifier

RCBT tries to reduce the chance of classifying test data with default class by building a series of stand-by classifiers apart from the main classifier. Moreover, RCBT carefully combines a subset of lower bound rules to make a collective decision instead of selecting only one shortest lower bound rule as CBA does. The subset of lower bound rules are selected based on the discriminate ability of genes. In this way, RCBT will not miss globally significantly rules which are



unable to be identified because of advance feature selection, while concentrate on a small number of informative genes.

5.3 Building Classifier

Let RG_j denote the *set* of rules groups, each of which is a top-j rule group for at least one training data of a certain class. Will thus have k sets of rule groups RG_1, RG_2, ..., RG_k. These k sets of rule groups are used to build k classifiers CL_1, CL_2, ...,CL_k with CL_j being built from RG_j. Call CL_1 the main classifier and CL_2, ..., CL_k backup classifiers. For each rule group in RG_j, RCBT finds its nl shortest lower bound rules by calling algorithm FindLB().

Dataset	RCBT	CBA	IRG Classifier	C4.5 family			SVM
				single tree	bagging	boosting	
AML/ALL (ALL)	91.18%	91.18%	64.71%	91.13%	91.18%	91.18%	97.06%
Lung Cancer(LC)	97.99%	81.88%	39.93%	81.83%	96.64%	81.88%	96.64%
Ovarian	97.67%	93.02%	-	97.67%	97.67%	97.67%	97.67%
Cancer(OC)	97.06%	82.35%	38.24%	26.47%	26.47%	26.47%	79.41%
Prostate Cancer(PC)							
Average Accuracy	95.93%	87.11%	80.95%	74.3%	77.99%	74.3%	92.70%

Table 5.1: Classification Results

VI CURLER: FINDING AND VISUALIZING NONLINEAR CORRELATION CLUSTERS

Detecting nonlinear correlation clusters is challenging because the clusters can have both **local** and **global** orientations, depending on the size of the neighborhood being considered. As an example, consider Figure 6.1, which shows a 2D sinusoidal curve oriented at 45 degrees to the two axes. Assuming the objects cluster around the curve, will be able to detect the global orientation of this cluster if consider a large neighborhood which is represented by the large circle centered at point *p*. However, if take a smaller neighborhood at point *q*, will only find The local orientation which can be very different from the global one. Furthermore, the local orientations of two points that are spatially close may not be similar at the same time, as can be seen from the small neighborhoods around *q* and *r*.

6.1 Algorithm

EM Clustering: A modified expectation-maximization subroutine *EM Cluster* is applied to convert the original dataset into a sufficiently large number of refined micro clusters with varying orientations. Each microcluster M_i is represented by its mean value μ_i and covariance matrix Σ_i.

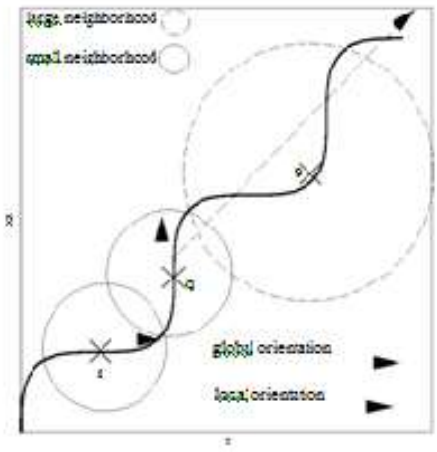


Figure 6.1: Global vs Local Orientation

At the same time, a similarity measure called co-sharing level between each pair of microclusters is computed.

Cluster Expansion: Based on the co-sharing level between the microclusters, a traversal through the microclusters is carried out by repeatedly choosing the nearest microcluster in the co-shared – *neighborhood* of a currently processed cluster. Denote this subroutine as *ExpandCluster*.

NNCO plot (Nearest Neighbor Co-sharing Level & Orientation plot): In this step, nearest neighbor co-sharing levels and orientations of the microclusters are visualized in cluster expansion order. This allows us to visually observe the nonlinear correlation cluster structure and the orientations of the micro clusters from the NNCO plot.

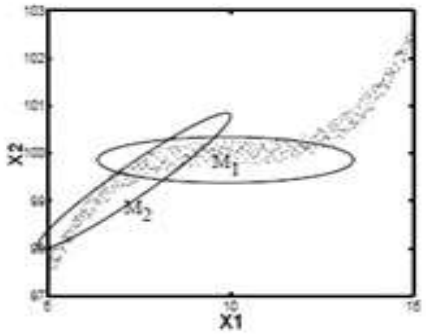


Figure 6.2 : Co-sharing between Two Microclusters

6.2 Top-down clustering

Having identified interesting clusters from the orientation plot, it is possible to perform another round of clustering by focusing on each individual cluster. The reason for doing so is that the orientation captured by the initial orientation plot could only represent the global orientation of the clusters.Each data object is assumed to have membership probabilities for several micro clusters in CURLER. Define the **data members** represented by a discovered cluster C which consists of micro cluster set MCS as the set of data objects whose highest membership probabilities are achieved in the micro cluster .

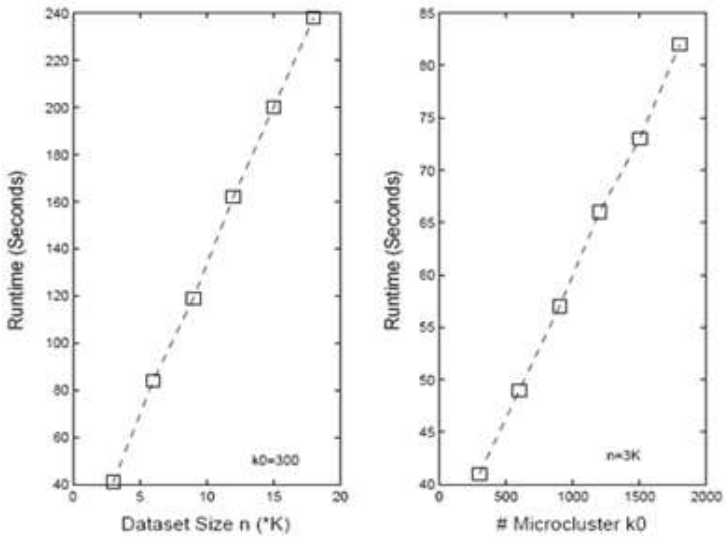


Figure6.3: Runtime vs Dataset Size n and # Microclusters k_0 on the 9D Synthetic Dataset

6.3 Synthetic Dataset

The difficulty of getting a public high-dimensional dataset of well-known nonlinear cluster structures, compared the effectiveness of CURLER with 4C on a 9D synthetic dataset of three helix clusters. The three helix clusters existed in dimensions $1-3$ (cluster 1), $4-6$ (cluster 2), and $7-9$ (cluster 3) respectively and the remaining six dimensions of each cluster were occupied with large random noise, approximately five times the data. Each cluster mapped a different color: red for cluster 1, blue for cluster 2, and yellow for cluster 3, as shown in Figure 6.4. Below is the basic generation function of helix, where $t \in [0, 6\pi]$,

$x_1 = c * t,$

$x_2 = r * sin(t),$

$x_3 = r * cos(t).$

The top-level NNC plot in Figure 6.2 shows that all the three clusters were identified by CURLER in the sequence of cluster 1, cluster 3 and cluster 2, separated

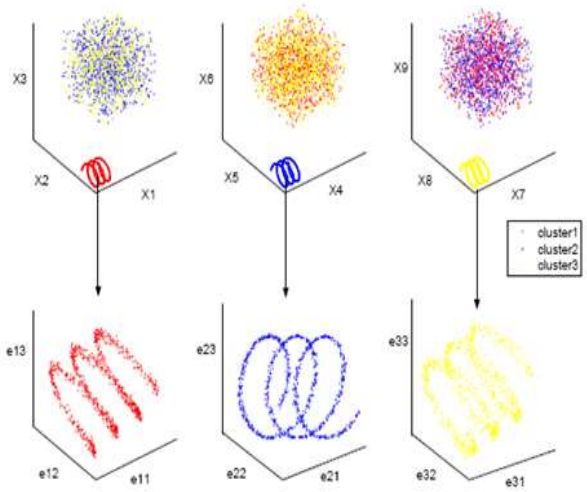


Figure 6.4 : Projected Views of Synthetic Data in both Original Space and Transformed Clustering Spaces

by two NNC-zero-gaps. The top-level orientation plot further indicates the cluster existence subspace of each cluster, the gray dimensions. The noise dimensions are marked with irregular dazzling darkening and brightening patterns.

6.4 Real Case Studies

To have a rough idea of the potential of CURLER in practical applications, applied the algorithm to three real-life datasets in various domains. Our experiments on the iris plant dataset, the image segmentation dataset, and the Iyer time series gene expression dataset show that CURLER is effective for discovering both nonlinear and linear correlation clusters on all the datasets above. As the cluster structures of the first two public datasets have not been described, will begin our discussion with the examination of their data distributions with the projected views. Will only report the top-level clustering results of CURLER here due to space constraint.

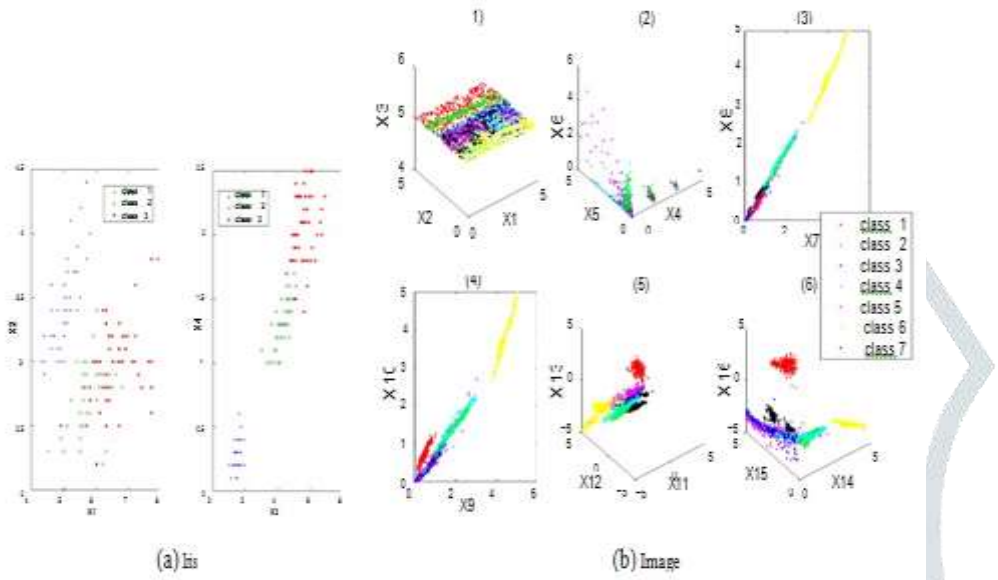


Figure 6.5: Projected Views

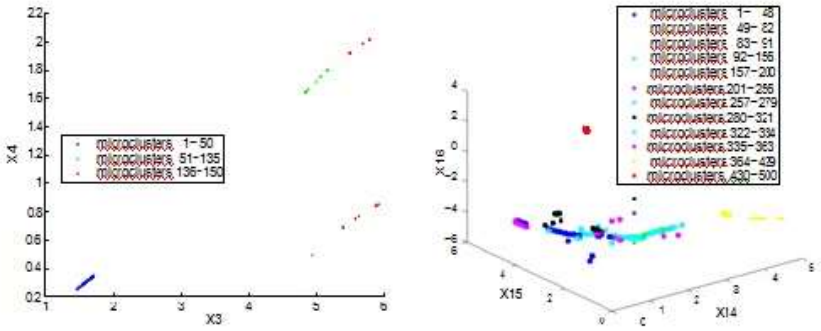


Figure 6.6 (a) Iris (b) Image

VII IMPLEMENTATION

Gene expression clustering algorithms may be classified into two big categories: full space clustering algorithms which evaluate the expression profile similarity of genes in all conditions, and *subspace clustering* algorithms which evaluate similarity in a subset of conditions. The most commonly applied full space clustering algorithms on gene expression profiles are hierarchical clustering algorithms, self-organizing maps, and K-means clustering algorithms, Hierarchical algorithms merge genes with the most similar expression profiles iteratively in a bottom-up manner. Self organizing maps and K-means algorithms partition genes into user-specified *k* optimal clusters. Other full space clustering algorithms applied on gene expression data include Bayesian network and neural network. Density-based subspace clustering algorithms, and our CURLER algorithm too, would assign each data object (gene) to just one cluster. Bi clustering algorithms provide an answer to this problem which allow overlapping clusters. These algorithms require genes of the same cluster to be dense and close to each other in correlated subspace.

7.1 Regulation Measurement

Suppose d_{ic_a} and d_{ic_b} are the expression levels of gene g_i under conditions c_a and c_b respectively. Could then say g_i is **up-regulated** from condition c_b to condition c_a, denoted as $Reg(i, c_a, c_b) = Up$, if the increase in expression level exceeds its regulation threshold γ_i, as described in Equation 5.3. Alternatively, say g_i is **down-regulated** from condition c_a to c_b, denoted as $Reg(i, c_b, c_a) = Down$. In this case, call c_b the **regulation predecessor** of c_a, denoted as c_b $\mathcal{X}$ c_a, and c_a as the **regulation successor** of c_b for g_i, denoted as c_a $\mathcal{Y}$ c_b (the arrow always points from bigger value to smaller value). Otherwise there is no regulation between c_a and c_b for g_i.

$$Reg(i, c_a, c_b) = \begin{cases} Up & \text{if } d_{ic_a} - d_{ic_b} > \gamma_i \\ Down & \text{if } d_{ic_a} - d_{ic_b} < \gamma_i \end{cases}$$

In this chapter, for ease of understanding, assume the regulation threshold of g_i, γ_i, as a pre-defined percentage of the expression range of g_i in Equation5.4, where n is the dimensionality of the expression dataset and γ is a user-defined parameter ranging from 0 to 1.0. Consider imposing a regulation threshold important for pattern validation, as it will help to distinguish useful patterns from noise. In practice, other regulation thresholds, such as the average difference between every pair of conditions whose values are closest, normalized threshold , average expression value , etc., can be used where appropriate.

$$\gamma_i = \gamma \times (MAX_{1 \le j \le n}(d_{ic_j}) - MIN_{1 \le j \le n}(d_{ic_j}))$$

The intuition behind using a local regulation threshold for different genes instead of a global one is that individual genes have different sensitivities to environmental stimulations. For instance, studies in reveal that the magnitudes of the rise or fall in the expression levels of a group of genes inducible or repressible by hormone E2 can differ by several orders of magnitude.Current pattern-based and tendency-based models can only cope with the extreme and probably biased case where $\gamma = 0$, and is constrained to the positive correlation. If $\gamma > 0$, these models become problematic.

To support this general concept of regulation, a naive approach is to record the regulation relationships between all possible pairs of C^2 conditions. Instead, propose a new model, called $RW\ ave^{\gamma}\ ^1$, which only keeps the regulation information of *bordering condition-pairs* for the genes in a wave-boosting manner with respect to γ. Figure 5.3 illustrates the $RW\ ave^{0.15}$ model ($\gamma_1 = \gamma_2 = 4.5$ and $\gamma_3 = 1.8$) for the running example (Table 5.1). $c_5 - c_1$ is one bordering condition-pair for g_1 since it represents the smallest interval above $\gamma_1 = 4.5$. Consequently, any condition c_i that lies on the left hand side of c_5 will guarantee to have a bigger difference than γ_1 when compared to any condition c_j that lies on the right hand side of c_1.
as can be seen, there is no need to keep the regulation information of non-bordering pairs. The formal definition of the RW aveγ model is given below.

7.2 Effectiveness

Ran the reg-cluster algorithm on the 2D 2884 × 17 yeast dataset with $MinG = 20$, $MinC = 6$, $\gamma = 0.05$ and $\varepsilon = 1.0$; 21 bi-reg-clusters are output in 2.5 seconds, where the overlapping percentage a bi-reg-cluster with another one generally ranges from 0%

to 85%. Note that did not perform any splitting and merging of clusters. Due to space limit, only report the details of three non-overlapping bi-reg-clusters with 21 genes and six conditions each.

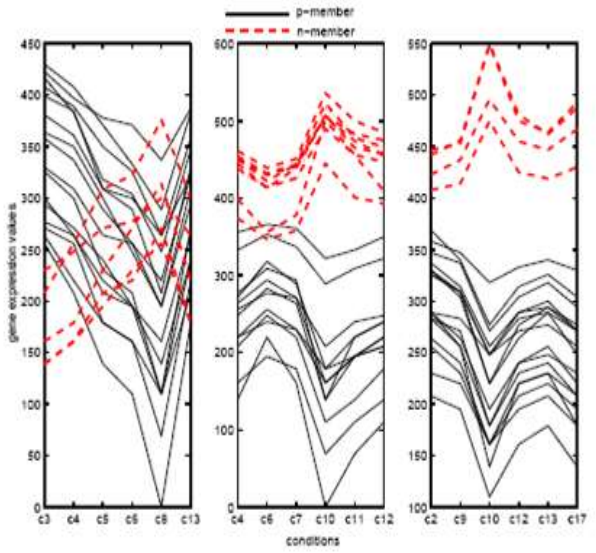


Figure 7.1 : Three biclusters

Figure 7.1 illustrates the gene expression profiles for each of the three bi-reg clusters. Our reg-cluster algorithm can successfully identify shifting-and-scaling patterns satisfying the regulation and coherence thresholds, where the scaling factor can be either positive or negative. For each bi-reg-cluster, represent its p-members with black solid lines and its n-members with red dashed lines. Obviously, the relationship between any two p-member genes or between any two n-member genes of the same cluster is shifting-and-positive-scaling while that between a p-member gene and a n-member gene is shifting-and-negative-scaling. As a remarkable characteristic of reg-clusters, crossovers can be observed frequently in the gene expression profiles of a pair of genes, resulting from the combination effects of shifting and scaling. In contrast, previous pattern-based biclustering algorithms only allow pure shifting or pure positive-scaling patterns (but not a mixture of both) and hence fail to identify the three bi-reg-clusters.

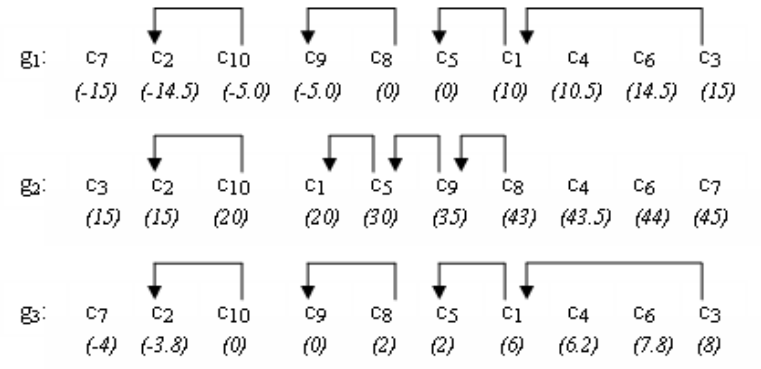


Figure 7.2 $RW\ ave^{0.15}$ Models

VIII CONCLUSION

In recent years, large amounts of high-dimensional data, such as images, handwriting and gene expression profiles, have been generated. Analyzing and handling such kinds of data have become an issue of keen interest. Elucidating the patterns hidden in high-dimensional data imposes an even greater challenge on cluster analysis. It is proposed effective and efficient data

mining methods for gene expression analysis in capturing the correlation between gene expression profiles and environmental conditions, and also the correlation among genes themselves. While focus on gene expression data, our data mining techniques can be applied to other kinds of high-dimensional data with homologous correlations as well.

The high-dimensionality of gene expression data renders traditional item wise association rule mining algorithms impractical due to exponential explosion of item combinations. Although a recent row wise rule mining algorithm FARMER is much more efficient than traditional item-wise algorithms by identifying interesting rule groups instead of searching individual rules one by one, the number of interesting rule groups can still be very large. Proposed the concept of top k covering rule groups, Top KRGs, and developed an efficient algorithm for Top KRGs discovery. In this way, not only solved the problems of inefficiency and huge rule number, but also helped users concentrate on the most significant information and minimized the information loss .Experimental studies on four benchmark gene expression datasets demonstrate that our Top KRGs algorithm is significantly faster than FARMER.

Based on Top KRGs, designed a novel associative classifier RCBT composed of a committee of *k* sub-classifiers. Each test sample is classified by the highest ranked sub-classifier and will be assigned the default class only when no sub-classifiers matches the test sample. Compared with previous associative classifiers, RCBT greatly reduces the chance of default class judgment as well as successfully locating globally significant rules. Moreover, by combining the discriminating powers of the delicately selected rules from Top-KRGs, RCBT achieves a rather high classification accuracy on four benchmark gene expression datasets. To give users some hints on Top-KRGs criteria, effective visualization techniques are also introduced, which provides an interactive graphic interface for users to observe, compare and explore rule groups.

To address nonlinear correlation, proposed a novel algorithm CURLER which adopts a fuzzy EM clustering subroutine to estimate the nonlinear orientations of the data in a trade off for efficiency and accuracy. Inspired by the reachability plot of OPTICS, it also proposed NNCO plot which visualizes the clusters embedded in subspace as well as their orientations. As another contribution, CURLER works in top-down manner so that users are able to further explore the sub-structure of any cluster of their interest. Experimental studies were carried out on synthetic helix datasets, UCI machine learning repository and real-life gene expression data to show the efficiency and effectiveness.

REFERENCES

1. Agrawal, R., Imielinski, T. and Swami, A. ―Mining associations between sets of items in massive databases.‖ ACM SIGMOD International Conference on Management of Data, pp 207--216, Washington, DC, May 1993.

2. Hipp, Jochen , Guntzer, Ullrich and Nakhaeizadeh, Gholamreza, ―Algorithms for Association Rule Mining – A general Survey and Comparison‖. SIGKDD explorations, Vol 2, Issue – 1, pp 58 – 63, Mar – 2004.

3.Goldberg, David E. Genetic Algorithms in Search, Optimization and Machine Learning.

4. Boston: Addison-Wesley Longman Publishing Co., 1989

5.Punch W. F., Pei M., et al ―Further Research on Future Selection and Classification using Genetic Algorithms‖, 5th International Conference on Genetic Algorithm, Champaign IL, pp 557 – 564, 1993.

6.Luscombe, N.M., Greenbaum, D. and Gerstein, M : ―What is Bioinformatics? A Proposed Definition and Overview of the Field.‖.Methods of Information in Medicine,40(4),pp 346-358,May 2001

7.. Piatetsky-Shapiro, G. and Tamayo, P : ―Microarray Data Mining: Facing the Challenges.‖ SIGKDD xplorations, 5(2), pp 1-5, 2003 .

8.. Liu, L., Yang, Jiong. and Tung, Anthony. ―Data Mining Techniques.

9. Microarray Datasets.‖ Proceedings of the 21st International Conference on Data 13

10. Engineering 2005 IEEE. pp 182-192 ,2005

11. Shah, Shital C. and Kusiak, Andrew, ― Data Mining and Genetic Algorithm Based Gene Selection‖, Artificial Intelligence in Medicine 2004, (31), pp 183-196, Vol – 2139, 2004.