Bottom up Biclustering Using One Leaveout Correlation

By J. Bagyamani, K. Thangavel, R. Rathipriya

Periyar University, Salem, TamilNadu, India

Abstracts - Uncovering genetic pathways are equivalent to finding clusters of genes with expression levels that evolve coherently under subsets of conditions. Analysis of gene expression data is used in many areas including drug discovery and clinical applications. Biclustering, a relatively new unsupervised learning technique, allows the assignment of individual genes to multiple clusters. It can help researchers discover useful information related to the function of genes. This proposed MAXBIC+ biclustering technique uses bottom up approach to extract two biclusters at a time with different objectives based on the query gene. It includes three steps viz. gene selection, novel method of identifying initial bicluster seed with reference to the query gene and growing the seed based on greedy approach till an optimal bicluster is obtained. Experimental results show that the proposed novel approach is effective in finding interesting biclusters. Since this algorithm uses Average Spearman Ratio (ASR) as merit function during growing phase, it is able to discover coherent biclusters, which could not have been identified with the use of the Mean Squared Residue (MSR) score. Experiments conducted on two benchmark datasets effectively find the biclusters with significant Gene Ontology (GO) based functional meaning.

Keywords: Biclustering, Gene Expression data, Gene Selection, Bottom-Up approach, Gene Ontology.

Classification: GJCST H.3.3
Bottom up Biclustering Using One Leaveout Correlation

J.Bagyamani¹, K. Thangavel², R. Rathipriya²

Abstract: Uncovering genetic pathways are equivalent to finding clusters of genes with expression levels that evolve coherently under subsets of conditions. Analysis of gene expression data is used in many areas including drug discovery and clinical applications. Biclustering, a relatively new unsupervised learning technique, allows the assignment of individual genes to multiple clusters. It can help researchers discover useful information related to the function of genes. This proposed MAXBIC+ biclustering technique uses bottom up approach to extract two biclusters at a time with different objectives based on the query gene. It includes three steps viz. gene selection, novel method of identifying initial bicluster seed with reference to the query gene and growing the seed based on greedy approach till an optimal bicluster is obtained. Experimental results show that the proposed novel approach is effective in finding interesting biclusters. Since this algorithm uses Average Spearman Ratio (ASR) as merit function during growing phase, it is able to discover coherent biclusters, which could not have been identified with the use of the Mean Squared Residue (MSR) score. Experiments conducted on two benchmark datasets effectively find the biclusters with significant Gene Ontology (GO) based functional meaning.

Keywords: Biclustering, Gene Expression data, Gene Selection, Bottom-Up approach, Gene Ontology.

I. INTRODUCTION

A gene expression matrix is made up of many genes all examined under a set of conditions or samples. According to Madeira S C and Oliveira A L (2004) biclustering is a data mining technique which allows simultaneous clustering of the rows and columns of a matrix. Gene expression analysis can be done by applying a biclustering procedure to gene expression data. A gene may belong to no cluster, one or more clusters. According to Bryan K., Cunningham, P (2006), top-down methods tend to produce a more uniform set of clusters in terms of size and shape. This representation may not accurately model the diverse set of functional modules that may be present in expression data. Also because they deal initially with the full set of dimensions, top-down approaches may not scale well as the dataset increases in size. Many approaches, which are based on this homogeneity, function follow Bottom-up strategy. This paper is organized as follows: Section 2 details the preliminary of gene expression data along with literature survey and the evaluation measures. Section 3 explains the proposed work. Section 4 provides the experimental results and Section 5 concludes the paper.

II. BACKGROUND

The earliest biclustering algorithm is direct clustering by Hartigan (1972) also known as block clustering. The original Cheng and Church (2000) node deletion algorithm began with the entire dataset and iteratively deleted rows and columns gradually minimizing the Mean Squared Residue score of the sub-matrix. In Liu and Wang (2007), the Maximum Similarity Bicluster is extracted respect to the ’reference gene’. Bagyamani(2010a) maximum similarity bicluster is identified using multiple node deletion algorithm and evaluated with ACV (Average Correlation Value). These algorithms have used Top Down approach which has its own drawbacks. Ayadi, et.al (2009) defined a new measure namely Average Spearman Ratio (ASR) for identifying a bicluster. Though this algorithm is efficient they neglect to evaluate their bicluster models from a biological perspective by assessing the functional relationships of the genes in the biclusters. This proposed method identified the following drawback in Ayadi, et.al (2009) in finding the single-gene biclusters. The preprocessing method is applied by them to find single-gene biclusters. Though this preprocessing technique is used in Wang (2007) and Chun Tang (2001) for normalizing the data, the single-gene biclusters obtained using this normalisation has no significance. Thus preprocessing technique cannot determine the single-gene bicluster efficiently. In this proposed algorithm, a novel method namely ’one leave-out correlation’ has been used for identifying a single-gene biclusters. This proposed query based approach is significant since it discovers two coherent biclusters with reference to the query gene viz, coherent biclusters with high ASR and coherent biclusters with high volume.

1) Problem specification

A gene expression database can be regarded as consisting of three parts – the gene expression data matrix, gene annotation and sample / condition annotation. Let G be a set of genes, C a set of

About°: Department of Computer Science, Government Arts College, Dharmapuri - 636705, TamilNadu, India. Email: bagyagac@gmail.com
About°: Department of computer Science, Periyar University, Salem, TamilNadu, India. Email: drktvelu@yahoo.com, rathipriyar@yahoo.co.in

©2011 Global Journals Inc. (US)
conditions, and $E(G,C)$ the expression matrix of order $m \times n$. The aim of biclustering is to extract the sub-matrix $E(G', C')$ where $G' \subset G$ and $C' \subset C$ meeting some criteria with reference to query gene. The size or volume of the bicluster is defined as $|G'| \times |C'|$ where $|.|$ represents number of elements.

2) Bicluster Measures

Cheng and Church (2000) used MSR to evaluate the biclusters. One notable drawback of the MSR score is that it is also affected by variance, favoring correlations with low variances. However, biclusters should contain subset of genes showing similar behavior, not similar values. Ayadi et.al. (2009), introduced a new evaluation function called Average Spearman's Rho (ASR) which is used as a merit function in this proposed algorithm. ASR measure is similar to ACV but Spearman correlation is used. Another statistical score for evaluating a bicluster is the Average Correlation Value (ACV) defined by Teng L., Chan L. (2007)

\[
ACV = \max \left\{ \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{|c_{rowij}|}{m^2 - m}, \sum_{p=1}^{n} \sum_{q=1}^{n} \frac{|c_{colpq}|}{n^2 - n} \right\}
\]  

(1)

in which $c_{rowij}$ is the Pearson’s correlation coefficient between rows ‘i’ and ‘j’ and $c_{colpq}$ is the Pearson’s correlation coefficient between columns ‘p’ and ‘q’. Bicluster with high ACV (i.e., ACV approaching 1) is a good bicluster. Though both measures ACV and ASR are based on correlation, ACV uses Pearson correlation and ASR uses Spearman correlation.

3) Choice of Merit function

The choice of merit function is strongly related with the characteristics of the biclusters each algorithm aims to find. There is no uniform measure for coherence. Hartigan JA (1972) used variance measure; Shyama Das and Sumam Mary Idicula(2010) used $H$ score; Bryan, K., Cunningham, P (2006) used $H$ score; Cheng K et.al(2008) and Bagyamani J(2010a), used ACV measure; Ayadi et.al., (2009) and Bagyamani J (2010b) used ASR measure. These are the few coherence measures available in literature. These coherence measures evaluate the resultant bicluster. Since most biclustering algorithms use the coherent measures as merit functions namely Ayadi et.al., (2009) or as fitness function by Shyama Das and Sumam Mary Idicula(2010) the choice of coherence measure plays vital role in deciding the quality of the bicluster. From the Table 1, ACV and ASR identifies more types of biclusters than variance related measures. When the two variables being compared are monotonically related, even if their relationship is not linear the Spearman correlation is 1. In contrast, this does not give a perfect Pearson correlation. Hence ASR is a good measure to evaluate a bicluster than the ACV. Hence in this proposed work ASR is used as merit function to grow biclusters. Since we apply bottom up approach to find biclusters, the efficient method of finding the initial biclusters play a vital role and the quality of the resultant bicluster depends on the initial biclusters.

### Table 1. Nature of Coherence Measures

<table>
<thead>
<tr>
<th>Sl.No</th>
<th>Coherence Measure</th>
<th>Nature of bicluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Variance</td>
<td>Constant</td>
</tr>
<tr>
<td>2</td>
<td>MSR or $H$-score</td>
<td>Constant, Constant row / column</td>
</tr>
<tr>
<td>3</td>
<td>$H$-score</td>
<td>Constant, Constant row / column, Additive coherent</td>
</tr>
<tr>
<td>4</td>
<td>ACV</td>
<td>Constant, Constant row / column, Additive coherent, Multiplicative coherent</td>
</tr>
<tr>
<td>5</td>
<td>ASR</td>
<td>Constant, Constant row / column, Additive coherent, Multiplicative coherent, Coherent evolution (up or down regulated)</td>
</tr>
</tbody>
</table>

III. PROPOSED WORK

Among large number of genes only small part of the genes are functionally important. In this paper, Roy Varshavsky et. al., (2006) used filter approach in which features are scored individually based on t-test. Insignificant genes are removed in this step. For the yeast gene expression dataset, t- scores of the genes lie between 0 and 2.7265 and the genes whose t-score > mean (t-score) are selected. Thus 2261 genes among 2884 genes are selected in the gene selection phase. Selecting a query gene is the foremost step in this proposed algorithm. A gene from selected genes with more functionality is set as query gene.

1) **Formation of single-gene biclusters using one leave-out correlation**

Most measures often gave high scores to dissimilar curves. Some pair of genes that are dissimilar receives a high score from the similarity measure, as a false positive. Correlation coefficient performs better than other measures, still results in many false positives. Laurie J et.al.,(2006) identified the following drawback of using correlation. The correlation between the two genes RFC2 (YJR068W) and NMD5 (YJR132W) is evaluated as 0.87 at all time points. But when the time point $t = 100$ is...
removed the correlation between the same genes is -0.29 (i.e., negatively correlated with very low correlation). The single-gene biclusters with reference to the query gene are identified using one leave out correlation. Let gq be the reference gene and gi are the other genes and t = (t1, t2, ... tn) are the time points then the correlation between these two genes by leaving the time point t = ti is computed by finding the Pearson’s correlation between gq (t1, t2, t3, ..., ti-1, ti+1...tn) and gi (t1, t2, t3, ..., ti-1, ti+1...tn). Those time points for which the correlation between the reference gene and the other genes gi that are higher than the threshold δ are removed i.e., made null entries. Thus each gene is expressed only in a subset of conditions. Each gene and the corresponding ‘condition subset’ thus constitute single-gene biclusters.

2) Seed formation phase

Next step aims to identify the local optimal sub-matrix as seed. The template / query gene gq, is a small window (gq, bq) of size (1, |bq|), where bq is the subset of conditions corresponding to gene ‘q’and |bq| is the number of conditions corresponding to gene ‘q’. Scan the single gene biclusters to generate sub matrices (gqUgi , bq∩bi) of size (2, |bq ∩bi|) where |bq ∩bi| is the number of conditions common to query gene ‘gq’ and gene gi. Then those sub matrices whose ASR is more than the threshold namely C1, C2, ...,Ck where k<m are selected. Let this collection be C. Among these sub matrices, one sub matrix with high ASR and high volume is taken as the valid seed. Let it be Cs. Other sub matrices or initial biclusters C - Cs are used for growing the seed. Here we apply anti monotone property i.e., if a sub matrix ‘L’ of size (2, |bq ∩bj |) for some gene gj has low ASR value, then any sub matrix formed by combining L with any other sub matrix will not lead to a bicluster with high ASR. So these sub matrices are pruned at the initial stage itself.

Schematic diagram of seed formation phase is given in Figure 1.

Fig.1. Bicluster seed formation phase

3) Bicluster Seed Growing Phase

In this phase the valid seed Cs= (giUgq, bi∩bq) is grown by combining the seed with other sub matrices Ci ∈ C- Cs to get intermediate seed. In order to arrive at a bicluster with maximum volume the valid initial biclusters obtained in the seed formation phase are arranged in descending order of their volume. Thus initial biclusters with more number of condition subset are combined with the seed first. Bicluster seed is grown by iteratively combining the valid seed with other sorted initial biclusters such that the intermediate seed should have ASR greater than the threshold. Thus the number of genes increase and number of conditions decrease during the growing phase.

Algorithm MAXBIC+

Input: E, the expression matrix.
Parameters: Threshold correlation δ, threshold ASR ϵ, Query gene gq, Minimum number of condition Mc.
Output: (i) a maximal bicluster with high ASR.  (ii) a maximal bicluster with high volume

Begin
1) ← Gs \ selected genes using t-score.
2) Select query gene gq ∈ Gs
3) (gi,Ti) = sg-bic (E(G,C),gq, δ) //find single-gene bicluster
4) C = init-bic(gi,Ti) // form init-bic as in Figure 1.
5) valid init-bic ← Prune init-bic with low ASR.
6) Sort C based on volume of valid seeds (Vs).
Select a seed $C_S$ from $C$ with high ASR, $C_S = \{initR, initC\}$; // vr(i) - rows in valid init-bic  
$i$//vc(i) - columns in valid init-bic $'i'$. 

Grow bicluster 
For $i=1$ to number of valid seeds $C_S = \{initR U Vr(i), initC N Vc(i)\}$ if (asr($C_S < \epsilon$) or (no. of conditions $< M_c$) 
Break 
Else grow bicluster with next valid init-bic. End 

Extract bicluster with maximum ASR. 
Extract bicluster with maximum volume.

$$function \ (g_i, T_i) = sg-bic \ (E(G,C), g_q, \delta)$$ 
for $g_i \in G$ 
$T_i = \{}$; 
for time point $t_i \in C$ 
Leave time point $t_i$ for $g_i$ 
Leave time point $t_i$ for $g_q$ 
$c(i,j) = \text{correlation}(g_i, g_q)$ 
if $c(i,j) < \delta$ 
$T_i = T_i U T_j$; 
End 
End 
End 

We set the minimum number of conditions in the bicluster to be $M_c$. This bicluster growing phase terminates until the combining process exhausts all the initial biclusters or the number of conditions in the bicluster is less than $M_c$. The intermediate biclusters are also pooled in order to arrive at a bicluster with maximal ASR and to arrive at a bicluster with maximal volume.

IV. EXPERIMENTAL ANALYSIS

In order to test the efficiency of the proposed algorithm the microarray datasets in table 3 have been considered for the experimental analysis.

Table 2. Microarray datasets

<table>
<thead>
<tr>
<th>Microarray Data</th>
<th>Size of original data</th>
<th>No. of genes after gene selection</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yeast Saccharomyces Cerevisiae</td>
<td>2884 x 17</td>
<td>2261</td>
</tr>
<tr>
<td>Colon cancer dataset</td>
<td>2000 x 62</td>
<td>1192</td>
</tr>
</tbody>
</table>

1) Comparison of biclusters 

Comparative study of the biclusters by SIMBIC and MAXBIC+ is given in Table 3. For the same reference gene, the size of the bicluster and ASR values of the biclusters of MAXBIC+ are high compared to biclusters of similar query based algorithm SIMBIC. Hence biclusters obtained by the proposed algorithm MAXBIC+ are efficient.

Table 3. Biclusters of SIMBIC Algorithm

<table>
<thead>
<tr>
<th>Query Gene</th>
<th>Size of bicluster</th>
<th>ASR</th>
<th>ACV</th>
</tr>
</thead>
<tbody>
<tr>
<td>288</td>
<td>19 x 14</td>
<td>0.8677</td>
<td>0.9224</td>
</tr>
<tr>
<td>133</td>
<td>20 x 15</td>
<td>0.8261</td>
<td>0.8319</td>
</tr>
<tr>
<td>374</td>
<td>27 x 16</td>
<td>0.8868</td>
<td>0.9105</td>
</tr>
</tbody>
</table>

Table 4. Biclusters of MAXBIC+ Algorithm

<table>
<thead>
<tr>
<th>Query Gene</th>
<th>Size of bicluster</th>
<th>ASR</th>
<th>ACV</th>
</tr>
</thead>
<tbody>
<tr>
<td>288</td>
<td>71 x 11</td>
<td>0.9804</td>
<td>0.9871</td>
</tr>
<tr>
<td>133</td>
<td>82 x 10</td>
<td>0.9810</td>
<td>0.9431</td>
</tr>
<tr>
<td>374</td>
<td>70 x 9</td>
<td>0.9821</td>
<td>0.9442</td>
</tr>
</tbody>
</table>

The following figures illustrate the bicluster plots of MAXBIC+ algorithm. Figure 2 represents the bicluster plot with $g_i = 288$ of Colon Cancer Data with high ASR. It has the volume 1650 (330 x 50) with ASR = 0.9507. Figure 3 illustrates the bicluster plot of the same query gene $g_i = 288$ with maximal volume = 4554 and ASR = 0.9041.
Figure 4 shows the Bicluster plot of Yeast data with query gene = 374 (YCL057W) which has high ASR = 0.9799. The 30 genes involved in this bicluster are YGR049W, YER056C, YMR049C, YGR214W, YMR143W, YML073C, YPR113W, YLR025W, YBL039C, YBR206W, YMR202W, YBR263W, YML045W, YKL166C, YLR075W, YJL190C, YML094W, YMR050C, YCL031C, YLR448W, YKR094C, YLR388W, YML081W, YJR027W, YAL033W, YMR211W, YCL057W, YGL246C, YHR073W, and YJL147C.

2) Gene Ontology

$P$ value of a bicluster provides the biological significance of a bicluster. It provides the probability of including genes of a given category in a cluster by chance. Thus overrepresented bicluster is a cluster of genes which is very unlikely to be obtained randomly. Suppose that we have a total population of $N$ genes, in which $M$ have a particular annotation. If we observe $x$ genes with that annotation, in a sample of $n$ genes, then we can calculate the probability of that observation, using the hyper geometric distribution. Thus the probability of getting $x$ or more genes with an annotation, out of $n$, given that $M$ in the population of $N$ have that annotation, is:

$$p_{value} = 1 - \sum_{j=0}^{x} \binom{M}{j} \binom{N-M}{n-j} \binom{N}{n}$$

The gene ontology namely Biological Process, Molecular Function and Cellular Component of the bicluster can be identified using GOTermfinder. In Figure 5, the biological network of the bicluster with high ASR with reference to the query gene 374 (YCL057W) is listed in Table 5. From the Tables 5, it is clear that the False Discovery Rate (FDR) is very low and in many occasions it is zero. Further the corresponding $p$ value is very small which show that there is very less probability to obtain the gene cluster in random.
Table 5: Gene Ontolgy – Biological Process of Yeast Data with (g_q=374)
The objective of this MAXBIC+ algorithm is to obtain two categories of biclusters with reference to the same query gene; one with maximal volume and the other with maximal ASR. Also the biclusters obtained in this proposed method has significant gene ontology with very low \( p \) value. An advantage of MAXBIC+ is that it is capable of identifying highly correlated biclusters with reference to the query gene. In future, a deep biological analysis of the intermediate biclusters obtained in this algorithm can be carried out to extract one optimal bicluster with high biological significance.

VI. ACKNOWLEDGEMENT

The first author acknowledges the UGC, SERO, Hyderabad to carry out this research under FIP. The second author acknowledges the UGC, New Delhi for financial assistance under major research project grant No. F-34-105/2008.

REFERENCES


This page is intentionally left blank