Discovery of Overlapping Pattern Biclusters from Gene Expression Data using Hash based PSO

Debahuti Mishra

*Institute of Technical Education and Research, Siksha O Anusandhan Deemed to be University, Bhubaneswar, Odisha, India

Abstract

Most of the biclustering algorithms believe on grouping the data elements on the basis of distance metric between objects and conditions, where as in real data, strong correlations may exist among a set of data elements and conditions even if they are far apart from the measuring candidate which can be identified in the form of similar patterns such as scaling and shifting patterns. Most of the biclustering methods usually perform their tasks under the assumption that each gene belongs to only one bicluster. But, depending on the experimental conditions being investigated, each gene may have similar expression pattern with different genes in different biclusters and they can, therefore, belong to more than one bicluster. In this paper, an efficient model has been proposed which captures the coherent behaviour among the data elements measuring the co-expression among data elements in the hybridized framework of hashing and Particle Swarm Optimization (PSO) and also discovers overlapping biclusters which can lead to discovery of the great biological complexity.

© 2011 Published by Elsevier Ltd. Selection and/or peer-review under responsibility of C3IT

Keywords: Biclustering; scaling pattern; shifting patterns; hashing; PSO

1. Introduction

Biclustering is a process of discovering biologically significant information from gene expression data [1-2]. Many traditional biclustering methods often assume that the biclusters rely on metric distance between objects and conditions. In fact, strong correlations may still exist among a set of objects even if they are far apart from each other as measured by the distance functions in the form of co-expressions among genes and conditions [3]. A group of co-expressed genes are the ones that demonstrate similar expression patterns over a substantial subset of samples. These pattern biclusters can be characterized in the form of shifting and scaling patterns [1-3] and is essential to discover the gene expression profiles which give better understanding of class discovery. Most of the biclustering methods usually perform their tasks under the assumption that each gene belongs to only one bicluster [4-7]. Several proteins perform multifunctional roles in a cell and they have to interact with different groups of proteins to fulfill them and also depending on the experimental conditions being investigated, each gene may have similar expression patterns with different genes in different biclusters and they can, therefore, belong to more than one biclusters [5]. In this paper, we concentrate on a new bicluster pattern based on correlations which overcomes the traditional biclustering model proposed by Cheng and Church [2] by discovering the significant pattern biclusters in the framework of hashing and PSO. This paper is organized as follows; section 2 deals with the study of related work; the experimental evaluation is given in section 3;
the result analysis is given in section 4 and in section 5 the conclusion and future scope of the work is provided.

2. Related Work

A model of discovering pattern based clusters by considering the scaling and shifting patterns formed by similar set of genes and conditions under pCluster model has been explored by Wang et.al [3]. Jian Pei et al. [8] proposed a model MaPle which conducts a depth-first, divide-and-conquer search strategy. MicroCluster model by Lizhuang Zhao et al. [9] have been proposed and used weighted, directed multi graph for finding the patterns formed by subset of genes and conditions. Wen-Hui Yang et al. [10] given an idea of mixed clustering algorithm. The model by Kin-On Cheng et al. [11] can identify both additive-related and multiplicative-related biclusters. Jiun-Rung Chen et al. [12] instead of generating MDS for every two genes, they have generated only MDS for every two conditions. To escape from local minima, evolutionary algorithms have been proposed to discover global optimal solutions. Junwan Liu et al. [13-15] proposed a novel multi-objective particle swarm optimization biclustering (MOPSOB) algorithm to mine coherent patterns from microarray data. Mohsen Lashkargir et al. [16] proposed a hybrid algorithm which is based on adaptive multi objective PSO for discovering biclusters in gene expression. A model for overlapping clustering (MOC) has been proposed by Banerjee et al. [7]. Cleuziou, G. et al. [17] proposed the Overlapping $k$-Means (OKM) based on a centre-based method that extended the $k$-means algorithm. Cano et al. [4] recently proposed a possibilistic spectral biclustering algorithm (PSB) to obtain potentially overlapping biclusters, based on fuzzy technology and spectral clustering. Rough overlapping biclustering (ROB) approach have been proposed by R.Wang et al. [5]. Patrik C. H. Ma et al. [6] proposed a novel information theoretic approach to discover overlapping clusters from gene expression data.

3. Experimental Evaluation

Our proposed model consists of four phases as shown in fig 1. For empirical comparisons; the synthetic data set of 200 objects and 10 conditions [18] and real data sets such as breast cancer of 699 objects and 10 conditions, leukemia of 7129 objects and 72 conditions and scaled expression consisting of 12488 objects and 12 conditions have been used [18-19]. In the phase-I proposed Rough-PCA method [20-21] applied to reduce the data sets.

![Gene Expression Data set](image)

Pre-processing using Rough-PCA [20-21]  
Discovery of initial seed patterns using hashing  
Discovery of pattern biclusters using PSO  
Discovery of overlapping pattern biclusters

Fig. 1. Proposed model for discovery of overlapping pattern biclusters in the proposed search strategy

Phase II: Discovering initial seeds [22]: In this phase, the difference of each object (gene) for all possible combinations of conditions, are found from the pre-processed dataset and sorted according to their differences called as difference matrix. After this, two matrices such as (1) sorted data matrix and (2) an index matrix that holds the index of object (gene) value of the data items from the sorted difference matrix is created. The set of genes that satisfy the condition pair have been discovered by searching for the gene index. The selected gene and condition pairs are kept into the pair matrix. Two another matrices such as, metaData and pointer matrix are created as shown in fig 2. metaData uses hash function to store the infoNode; where, infoNode contains two fields; first field stores the condition pair value and other field stores the index of another pointer matrix where details of the condition pair is stored. Pointer matrix is used to store the number of occurrences of condition pair and their index value in pair matrix. The index of hash table (metaData) varies from 0 to bucket size (user defined). The whole procedure is given in fig.2. As there are multiple entries in pair matrices, this is pruned and the new condition pairs are discovered. The object pair and condition pairs are used as initial seeds for the third phase to generate pattern biclusters.
Phase-III: Discovery of pattern based biclusters: In this phase; the initial population of gene and condition pairs as discovered in phase II are used as input for the PSO framework to mine pattern biclusters. After encoding the initial seeds, we get a binary matrix of swarms of \( \{ \text{size of initial swarm} \times (\text{genes} + \text{conditions}) \} \). Each bicluster is encoded as a particle of the population. Each particle is represented by a binary string of fixed length \( n \times m \), where \( n \) and \( m \) are the number of genes and conditions. The first \( n \) bits are associated to \( n \) genes, the following \( m \) bits to \( m \) conditions. If a bit is set to 1, it means that the responding gene or condition belongs to the encoded bicluster; otherwise it does not. The encoded biclusters are given in fig 3 as an example from synthetic data set.

Our approach is multi-objective in nature to mine biclusters with low mean squared residue, with high volume and row variance. The objectives are as follows (1):

\[
\begin{align*}
& f_1(x) = \frac{|G|}{|G| + |C|}, \quad f_2(x) = \frac{\text{MSR}(x)}{\delta} \quad \text{and} \\
& f_3(x) = 1 - \frac{\text{RVAR}(x)}{G \times C}
\end{align*}
\]

(1)

Where, \( G \) and \( C \) are the total number of genes and conditions of the microarray datasets respectively. \( \text{Size}(x) \), \( \text{MSR}(x) \) and \( \text{RVAR}(x) \) denotes size, mean squared residue (MSR) and row variance of bicluster encoded by the particle \( x \) and \( \delta \) is the MSR threshold. The discovered initial seeds, are treated as swarms \( \{X_1, \ldots, X_n\} \). Where, 1, ..., \( n \) is the number of genes paired with \( m \) number of conditions. In order to evaluate the degree of coherence MSR of the sub matrix \( A \) is [2] as follows:

\[
\text{MSR}(A) = \frac{1}{|I| \times |J|} \sum_{i \in I, j \in J} (a_{ij} - a_i - a_j + a_{ij})^2
\]

(2)

Where, \( I \) denotes the row set, \( J \) denotes the column set, \( a_{ij} \) denotes an element in the sub matrix, \( a_i \) denotes the \( i^{th} \) row mean, \( a_j \) denotes the \( j^{th} \) column mean, and \( a_{ij} \) denotes the mean of the whole bicluster. A sub matrix \( A \) is called a \( \delta \) bicluster if \( \text{MSR}(A) < \delta \) for some \( \delta > 0 \). \( \delta \) is the MSR threshold, high MSR value signifies that the data is uncorrelated. Our objective is to determine the swarms of largest size satisfying the fitness rule as given in (2). In each iteration, a new velocity value for each particle is calculated based on its current velocity, the distance from its previous best position and the distance from the global best position. The new velocity value is then used to calculate the next position of the particle is the search space. The particle’s velocity and the position are dynamically updated as follows:

\[
V_{id} = w \times V_{id} + c_1 \times \text{rand} \times (P_{id} - X_{id}) + c_2 \times \text{rand} \times (G_{id} - X_{id})
\]

(3)

The new velocity of a particle is updated by (3) taking into consideration of the particle’s previous velocity and previous position; \( w = [0.5 + \text{rand}/2] \) is an inertia weight and \( \text{rand} \) is a uniformly generated random number between 0 and 2. The values of various parameters are set to as; cognition parameter \( c_1=0 \), social parameter \( c_2=2 \), inertia weight \( \omega=0.5 \) and MSR value \( \delta = 0.007 \). \( P_{id} \) is the previous individual best position of this particle and \( G_{id} \) is the current global best position. Then (3) calculates the new position of the particle, \( X_{id} \) as follows:
\[ x_{id} = x_{id} + \delta_{id}. \text{If } \left( \text{rand}(\cdot) < S(\delta_{id}) \right) \text{ then } x_{id} = 1 \text{ else } x_{id} = 0.5(\delta_{id}) = \frac{1}{1 + \exp(-\delta_{id})} \quad (4) \]

The function \( S(v) \) is a logistic transformation as given in (4) and \( \text{rand}(\cdot) \) is a quasi-random number selected from a uniform distribution in \([0,1]\). For example, by applying the above method on fig 3 with we get the fig 4 shows the pattern biclusters discovered after application of PSO from synthetic data set. Termination condition can be set to maximum number of iteration (user defined) or the condition can be from where no new set can be derived. Before adding the new set into archive it is checked that the set is already present in the archive or not. If set is not present then it will be added or else rejected to avoid duplication.

**Phase-IV: Discovering overlapping biclusters:** In this phase, two rules rule1 and rule2 have been formed for discovering overlapping region. For given two biclusters, \( B_1 \) and \( B_2 \) having condition pair \( C_1 \) and \( C_2 \) with some gene set \( G_1 \) and \( G_2 \) respectively; there exists overlapping region between biclusters \( B_1 \) and \( B_2 \) if and only if: \( C_1 \cap C_2 \neq \Phi \) and \( G_1 \cap G_2 \neq \Phi \). Now, the percentages of participation (POP) of those set of genes which are common in both the condition pairs have been calculated using the following rules. Let \( G = G_1 \cap G_2 \) for biclusters \( B_1 \) and \( B_2 \).

- % of participation of \( G \) on \( G_1 \) (POP\(_1\)) = \((\text{Total element in } C_1 / (\text{Number of genes in } G \times \text{Number of condition pair in } C_1)) \times 100 \) (Rule 1)
- % of participation of \( G \) on \( G_2 \) (POP\(_2\)) = \((\text{Total element in } C_2 / (\text{Number of genes in } G \times \text{Number of condition pair in } C_2)) \times 100 \) (Rule 2)

This percentage of participation of genes can now be used for either removing all the overlapping modules from the biclusters or else it can be used to allow those overlapping modules to any one of the biclusters. If POP\(_1\), POP\(_2\) > \( \eta \), where \( \eta \) is a user defined threshold value (0.5), then we allow no change in the biclusters and overlapping among them could not be removed. Else if POP\(_1\, > \,\text{POP}_2\), then overlapping modules from bicluster \( B_2 \) is removed and there is no change in the elements of bicluster \( B_1 \), else overlapping modules from bicluster \( B_1 \) is removed without changing any elements in bicluster \( B_2 \).

### 4. Result Analysis

An extensive experimental study on the efficiency and effectiveness of proposed algorithm have been performed with synthetic and real-life data sets and our model is based on hashing technique applied in the PSO algorithm, it discovers the pattern biclusters with reasonable time and can be scaled as per the requirement of the data set by dynamically creating the space for new data elements. The proposed PSO have been compared with the models of SRPSO, KSRPSO and k-means [23]. Their convergence processes are recorded and compared and it has been found that the global best solution of proposed PSO becomes stable after 10 iterations whereas; other compared models are able to obtain the same global solution in more than 10 iterations as shown in fig 5. Finally, the overlapping modules among the biclusters or functional modules have been discovered and compared with some existing approaches like, rough overlapping model [5], value coherent model [17] and overlapping co-expression patterns [6] in different data sets given in fig 6 and our proposed model proves to be better with respect to number of overlapping modules discovered. Table 1 shows number of biclusters discovered run time (in seconds) and number of overlapping modules discovered using synthetic and real data sets.

<table>
<thead>
<tr>
<th>Data sets</th>
<th>No. of Biclusters</th>
<th>No. of overlapping biclusters</th>
<th>Run Time (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breast Cancer</td>
<td>118</td>
<td>33</td>
<td>90</td>
</tr>
<tr>
<td>Scaled Expression</td>
<td>2754</td>
<td>1375</td>
<td>740</td>
</tr>
<tr>
<td>Leukaemia</td>
<td>1647</td>
<td>6837</td>
<td>544</td>
</tr>
<tr>
<td>Synthetic</td>
<td>94</td>
<td>42</td>
<td>50</td>
</tr>
</tbody>
</table>

Fig. 5 Comparison of convergence process

Fig. 6 Performance analysis of proposed overlapping model with existing models
5. Conclusion and Future Work

This paper proposes a new pattern biclustering technique to discover pattern biclusters and also overlapping biclusters. We attempt to apply this biclustering algorithm to gene expression data matrix. Each gene in the subset of a bicluster is active or increasingly active over the set of conditions in the bicluster. Genes that seem to be working together under a set of conditions might possibly be linked to a similar function. When it comes to diseases that involve some sort of genetic disorder, biclustering gene expression matrices using samples of diseased tissues as well as healthy ones could lead to the discovery of genes that are active in the presence of the disease. As a future work, the biological significance of this model can be tested by measuring the $p$-value of the discovered biclusters and finding the biological grouping among the genes and conditions in the gene ontology, further as the proposed model is an unsupervised model, this can be extended to design a supervised or semi-supervised model.

References

18. Repository for Leukaemia and Scaled Expression datasets: www.broadinstitute.org