NEW CROSSING MINIMIZATION TECHNIQUE FOR CANCER DATASETS BICLUSTERING

Ahmed Sharaf Eldein
Helwan University
Faculty of Computers and Information
Information Systems Department
profase2000@yahoo.com

Samar Kassim
Ain Shams University
Faculty of Medicine
Medical Biochemistry and
Molecular Biology Department
samar_kassim@yahoo.com

Tamer Mohamed
Helwan University
Faculty of Computers and Information
Biomedical Informatics Department
Sniper_fci_cs@yahoo.com

ABSTRACT

Cancer is a fatal disease that causes large number of deaths all over the world; so many studies have been made on genes to study the behavior of this disease. Biclustering is one of the data mining techniques that aiming to find the behavior of subset of genes under subset of circumstances. In this paper we propose a new biclustering algorithm that uses the Barycenter value of each node and its position in the graph beside the weight and position of the neighbors of each node the graph to give a new weight for each node and reordering the nodes depending on this weight. The ordering of bipartite graph grouping the related nodes together which enhances the results of biclustering algorithm.

KEYWORDS

BaryCenter Heuristics, Biclustering, Gene Expression, Microarray, Leukemia.

1. INTRODUCTION

Cancer is a disease in which cells divide in uncontrolled way. This can occur because of damage in the DNA which cannot be repaired. New daughter cells will inherit the damaged DNA and proliferate very fast resulting in tumor formation. A tumor, or mass of cells, formed of these abnormal cells may remain within the tissue in which it is originated (a condition called in situ cancer), or it may begin to invade nearby tissues (a condition called invasive cancer). An invasive tumor is said to be malignant, and cells shed into the blood or lymph from a malignant tumor are likely to establish new tumors (metastases) throughout the body. Tumors threaten the life of individuals, when their growth disrupts the tissues and organs needed for survival [1].

DNA microarray is a technology that it is used to measure the amount of gene expression in the cell. It will help the Study of which genes are active and which are not at different situations. This will indicate the level of gene expression or the amount of gene products (proteins) expressed in the examined cells. This technology enables scientists to study what happens to genes in different disease conditions [2].

The result of microarray is a gene expression dataset which is a tow dimensional array - data table - where every row represents one gene and every column represent one condition or sample. Every cell in this table represent the expression level Lij of gene i in condition j, as in table 1.

There is many Data mining techniques used for analyzing gene expression datasets in order to extract information and finding relations between these data such as clustering and classification.

Clustering is used to assign data to groups or classes like classification but the difference is that in classification the classes are predefined but in clustering the classes are not determined before applying the algorithm. Clustering algorithms can be applied on gene expression datasets in order to group genes that have the same expression together and this is helping in the study of diseases such as cancer. There are three ways of clustering:

1. Gene-based clustering in this approach genes are treated as objects and samples are treated as features.

ISBN: 978-0-9853483-3-5 ©2013 SDIWC

2. Sample-based clustering is the opposite of gene-based clustering, samples are treated as objects and genes are treated as features.
3. Subspace clustering this technique combine the previous two approaches the genes and samples can be treated as objects and features so that gene may be object or features and samples are treated the same.

Biclustering or subspace clustering is the most efficient and more realty technique in clustering because clustering of genes depending on conditions or vice versa, and this is not happen exactly in the real life , a subset of genes may have similar behavior on subset of conditions, and here is the Biclustering come to solve this problem.

Gene expression datasets can be represented in bipartite graph where:

1. Genes are represented by the top layer.
2. Conditions are represented by the bottom layer.
3. Edges that connect the genes in the top layer with its corresponding conditions in the bottom layer.

The edge weigh wij is the expression value of gene i in condition j, where wij = 0 if there is no edge between node i and j and this means that the gene i does not express in condition j.

We can represent data in table 1 in a bipartite graph, as in figure 1,where top layer represents top conditions and bottom layer represents genes, then this Bipartite graph can be clustered directly but the results will be less accurate and in order to overcome this problem there is one more step before biclustering which is crossing minimization. Crossing minimization aimed to minimizing the number of crossing between lines that connect the two layers the graph is reordered and related nodes grouped together, as in figure 2, and this enhances the results of biclustering process.

Table 1. Gene Expression Data Matrix

	Cond 1	Cond 2	Cond 3	Cond 4
Gene 1	5	0	6	0
Gene 2	0	2	0	1
Gene 3	0	2	0	1
Gene 4	2	0	1	0
Gene 5	5	0	6	0

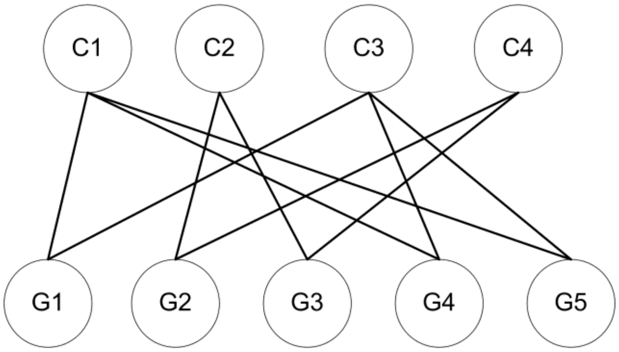

Figure 1. Bipartite graph representation of table 1

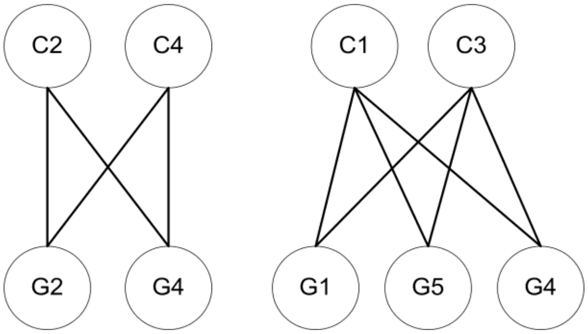

Figure 2. Bipartite Graph after crossing minimization

2. RELATED WORK

Biclustering algorithms can be classified in two main categories: graphical and non-graphical biclustering. We focus our discussions on graphical biclustering which is the main interest point of our research for general discussion on biclustering algorithms can be found in [3, 4]. There is another good reference that focuses on Barycenter crossing minimization which the main goal of our work is to enhance Barycenter algorithm [5].

Cheng and Church [6] define bicluster as a submatrix for which the mean square residue $H(I,J)$ of each bicluster is below than a predefined threshold. The algorithm is run in two main phases:

first, looping through the gene expression matrix removing rows and columns until $H(I,J)$ is less than the threshold; second, looping through the deleted rows and columns and adding to bicluster as long as $H(I,J)$is less than threshold. The $H(I,J)$of an element a_{ij} in a submatrix A_{IJ}

$$H(I,J) = \frac{1}{|I||J|}\sum_{i\in I, j\in J}(a_{ij} - a_{iJ} - a_{Ij} + a_{Jj})^2 \quad (1)$$

Where

Sub Row Average

$$a_{iJ} = \frac{1}{|J|}\sum_{j\in J} a_{ij} \quad (2)$$

Sub Column Average

$$a_{Ij} = \frac{1}{|I|}\sum_{i\in I} a_{ij} \quad (3)$$

Sub Matrix Average

$$\frac{1}{|I||J|}\sum_{i\in I, j\in J} a_{ij} \quad (4)$$

Ahmad and Ashfaq [7] extract clusters from dataset using bipartite crossing minimization biclustering techniques but instead of serial Bigraph crossing minimization using parallel algorithm in order to increase the performance of the algorithm beside the enhancement made on Cheng and Church biclustering algorithm to enable the local search for clusters instead of global search because after the Bigraph reordering the related values arranged together in blocks and global search is useless.

Ibrahim [8] using a Tabu search, which is a metaheuristics technique depending on guide local search technique to find global optimum solution. He adapts the Tabu search in Marti [9] to solve the crossing minimization problem and reach to better order of dataset. Tabu search run in two main steps: first, construction of initial solution and second, iterative improvement of this solution until reaching the optimum solution. Iterative improvement has two steps: Intensification and Diversification and each step of them has three steps: normal, influential, and opposite and the algorithm moves between these three steps as it insert or remove nodes depending on their Barycenter and if this move reduces the number of crossings in the graph and if there is no enhancement it stops after a specified number of iteration this process continue running in this way until reaching the optimum solution.

3. PROPOSED ALGORITHM

There are many techniques used for Biclustering one of these techniques is constructing a bipartite graph from gene expression dataset, reordering nodes in this graph using crossing minimization technique then apply biclustering algorithm on the reordered bipartite graph. Most of the algorithms used to crossing minimization in bipartite graph divided into three basic steps: the first step is the construction of bipartite graph, the second step is the initial ordering, and the last step is iterative improvement.

we calculate the rank of each node in layer 0 using Barycenter heuristics, as in equation 5, then reordering these nodes depending on this rank while keeping the layer 1 static then reordering layer 0 by calculating the rank of each node in layer 1 using equation 5, and reordering nodes while keeping the layer 0 static, and we will continue iteratively in this manner until no more changes occur in the order of nodes in the two layers.

$$Bary\ Center = \frac{\sum_{j\in Ni} W_{i,j} \times r_j}{\sum_{j\in Ni} W_{i,j}} \quad (5)$$

Where:
“Let v_i represent the i'th node in the non-static layer and set
N_i represents the set of neighbors of v_i. Also let r_j represent the rank of j'th member of the set N_i “[10].

We success to solve the problem of crossing minimization in a way that achieves better results in crossings number rather than Barycenter heuristics [11]. We use the same way of Barycenter heuristics used in [11] to minimize crossings in bipartite graph, but we make some changes to the calculations of weight of each node and using this weight as a new rank for reordering nodes. We calculate the Barycenter of each node as in algorithm 2, then calculating the new weight by adding the rank of each node to the sum of the ranks of its neighbors as in algorithm 3, and taking the position of each node into consideration of our calculations as in algorithm 4, then using this rank for reordering the nodes.

The algorithm of biclustering has three main basic steps:

- Construct bipartite graph
- Bipartite graph crossing minimization as in algorithm 1
- Applying biclustering algorithm on bipartite graph as in algorithm 5

Algorithm 1: Bipartite Crossing Minimization

```
1. Construct Bipartite graph BG
2. Compute the Barycenter for each
   node in the two layers, as in
   algorithm 2
3. Compute the new weight for each
   node in the two layers, as in
   algorithm 3
4. Reorder nodes in the two layers
5. Compute crossing minimization for
   bipartite graph
6. Repeat steps from 2 to 5 until
   the number of crossing  of the
   current bipartite graph is larger
   than the number of crossings of
   previous bipartite graph from
   last iteration
```

Algorithm 2: Computing the Barycenter for each node in the two layers

```
For all i such that $v_i$ is a vertex in top or bottom layer
Compute weighted mean for $v_i$
 IF Node$_i$.Rank != WeightedMean Then
  Node$_i$.Rank = WeightedMean
  PositionChanged = PositionChanged + 1
 End if
End for
```

Algorithm 3: Compute the new weight for each node in the two layers

```
FOR ALL i such that $v_i$ is a vertex in
top - or bottom - layer
$v_i$.NewRank  = | $v_i$.Rank |  + get $v_i$
neighbors Rank as in algorithm 4
END FOR
```

Algorithm 4: Get neighbors Rank of vertex

```
FOR Each neighbors j of vertex i
Sum + = (|$v_i$.Rank| - |neighbor$_j$.Rank|) *
(|$v_i$.Position| - |neighbor$_j$.Position|)
END FOR
```

After the process of crossing minimization we perform the Biclustering process as in algorithm 5, this algorithm find biclusters by performing local search because the fact that the crossing minimization process groups related rows and related columns together so the algorithm does not need global search to find the biclusters.

The algorithm uses mean square residue score - MSR - defined in equation 1 where MSR of each biclusters is less than a predefined threshold $^{\delta}$

Algorithm 5: Biclustering Algorithm

```
Comment: Identify Biclusters in Reordered
Matrix representation of Bigraph BG
Comment: Mean Squared Residue Score of Each
Bicluster is less than $^{\delta}$
Input: Bipartite graph after crossing
minimization
Output : Biclusters
  • startRow = 0;
  • startCol = 0;
  • Index = 0;
  • FOR i = 0 TO numCols DO
  • colSum[i] = 0;
  • blockSum[i] = 0;
  • Residue[i] = 0;
  • FOR r = startRow TO numRows DO
  • rowSum = 0;
  • FOR c = startCol TO numCols DO
      ▪ colSum[c]=colSum[c]+Matrix¯BG[r][c];
      ▪ rowSum = rowSum + Matrix¯BG[r][c];
      ▪ blockSum[c]=blockSum[c-1]+colSum[c];
      ▪ colCount = c - startCol + 1;
      ▪ rowCount = r - startRow + 1;
      ▪ blockCount = rowCount * colCount;
      ▪ rowMean = rowSum / colCount;
      ▪ colMean = colSum[c] / rowCount;
      ▪ blockMean = blockSum[c]/blockCount;
      ▪ Res = Matrix G[r][c] - rowMean -
        colMean + blockMean
      ▪ Residue[c] = Residue[c-1] + Res*Res;
      ▪ msr = Residue[c]/ blockCount;
      ▪ IF msr > $^{\delta}$ THEN
        -   Blocked = true;
        -   stopCol =  c;
        -   BREAK;
```

```
•   IF Blocked = TRUE THEN
•   Bicluster$_{bix}$ = Rows [startRow TO r],
    Columns[startCol TO  stopCol - 1];
•   Biclusters [Index] = bix;
•   Index = Index + 1;
•   startRow = r;
•   r = r - 1;
•   startCol = stopCol;
```

4. EXPERIMENTAL RESULTS

The proposed algorithm is written in C-sharp language and the experiments are running on i5 2.3 GHz processor and 4 GB of RAM, and under windows 7 operating system. We apply the algorithm on cancer gene expression datasets: lymphoma dataset [12], gastric cancer dataset [13], and AML prognosis dataset [14] to test the accuracy of the proposed algorithm relative to Barycenter used in SPHier algorithm. We apply the proposed algorithm and Barycenter algorithm on the three cancer datasets and compare the results of the two algorithms according the accuracy of crossing minimization.

After applying the proposed algorithm, and Barycenter algorithm of crossing minimization on the cancer datasets we get the results in table 2, that it is represented in figure 3.

Figure 3 shows enhancements that our algorithm made in the Barycenter algorithm, the proposed algorithm achieves better results especially in large size gene expression datasets with small number of conditions. Because the decrease in crossings number of bipartite graph leads to grouping related node together, the enhancement in crossings number increases the efficiency of biclustering algorithm [7].

Table 2. Crossings number of original dataset, Barycenter algorithm and proposed algorithm

	Original Dataset	Barycenter Algorithm	Proposed Algorithm
Gastric	11,048,714,141	11,048,713,498	11,048,654,635
AML Prognosis	131,451,569,724	131,451,501,126	131,450,516,864
Lymphoma	100,790,992,882	100,790,992,882	100,790,645,942

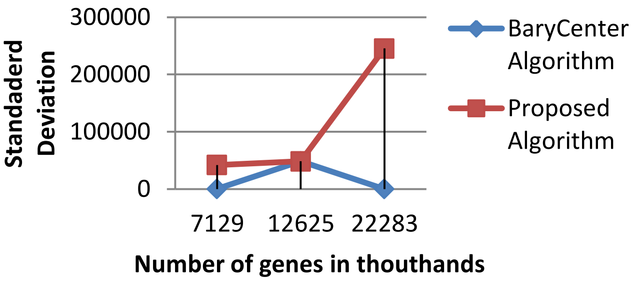

Figure 3. Comparison between standard deviation of Barycenter algorithm and proposed algorithm from the original crossings number

5. CONCLUSIONS AND FUTURE WORK

In this paper we enhanced Barycenter algorithm to refine the results crossing minimization of bipartite graph which will help in enhancing the results of bipartite graph biclustering algorithm.

In the proposed algorithm we add the rank of each node to the rank of its neighbors, and using the position of each node in the calculations, as we discussed in the proposed algorithm, to give a new rank to each node, which we depend on it to reorder the nodes in each layer.

when we analyze the results shown in figure 3 and by knowing that the number of genes and conditions of each dataset as stated in table 3, we conclude that the increase of number of conditions decreases the efficiency of our algorithm, and in case of small number of conditions we notice that the proposed algorithm gives more efficient results with large size gene expression dataset, as viewed in figure 3, the best result of gene expression datasets is the result of lymphoma dataset which has the minimum number of conditions and large size dataset, the proposed algorithm has a very important advantage which is in the worst case our algorithm gives the same result as Barycenter algorithm.

Table 3 number of genes and conditions of gene expression Datasets used in the experiment

	Number of genes	Number of conditions
Gastric	7,129	30
AML Prognosis	12,625	58
Lymphoma	22,283	29

We summarize the Future work for the proposed algorithm in the following directions: firstly, it can be enhanced to increase the efficiency of crossing minimization algorithm for minimum size gene expression datasets or for datasets with large number of conditions, and secondly, we can apply this algorithm in a parallel way to increase the speed of the algorithm because Barycenter algorithm does not depend on calculating the crossings number of bipartite graph to reorder the nodes while we use this way for reordering the nodes which takes more time in the calculation process.

6. REFERENCES

1. http://www.cancer.org

2. Macgregor, P.F. and Squire, J.A. 2002. "Application of microarrays to the analysis of gene expression in cancer". Clin.Chem. 48: 1170-1177.

3. Amos Tanay, Roded Sharany, and Ron Shamir, "Biclustering Algorithms: A Survey", 2004.

4. S. C. Madeira and A. L. Oliveira, "Biclustering algorithms for biological data analysis: A survey." IEEE/ACM Transactions on Computational Biology and Bioinformatics, 1, 2004.

5. A. Abdullah & A. Hussain, "A new Biclustering technique based on crossing minimization", Neurocomputing Journal 69, page 1882-1896, 2006.

6. Y. Cheng and G.M. Church. "Biclustering of expression data. In Proceedings of Intelligent Systems for Molecular Biology". 2000.

7. Waseem Ahmad, Ashfaq Khokhar, "SPHier: Scalable Parallel Biclustering Using Weighted Bigraph Crossing Minimization", 2007.

8. I. M. El Henawy, Ahmed Hussain Kamal and Ibrahim Hamed, "Using Tabu Search as a Crossing Minimization Technique toward Biclustering Gene Expression Datasets". In: Forth International Conference on Intelligent Computing and Information Systems, Cairo, Egypt, 2009.

9. Marti R. (1998), "A Tabu Search Algorithm for the Bipartite Drawing Problem", European Journal of Operational Research,Vol. 106, pp. 558-569.

10. Arifa Nisar, Waseem Ahmad, Wei-keng Liao, Alok N. Choudhary: "*High Performance Parallel/Distributed Biclustering Using Barycenter Heuristic*". SDM 2009: 1050-1062

11. Ahmad, W., Khokhar, A.: chawk: "A highly efficient biclustering algorithm using Bigraph crossing minimization". In: Second International Workshop on Data Mining and Bioinformatics, VDMB 2007, Vienna, Austria (Held in Conjunction with VLDB2007). (2007).

12. Raetz EA, Perkins SL, Bhojwani D, Smock K et al. Gene expression profiling reveals intrinsic differences between T-cell acute lymphoblastic leukemia and T-cell lymphoblastic lymphoma. Pediatr Blood Cancer 2006 Aug; 47(2):130-40.

13. Hippo Y, Taniguchi H, Tsutsumi S, Machida N et al. Global gene expression analysis of gastric cancer by oligonucleotide microarrays. Cancer Res 2002 Jan 1; 62(1):233-40.

14. Yagi T, Morimoto A, Eguchi M, Hibi S et al. Identification of a gene expression signature associated with pediatric AML prognosis. Blood 2003 Sep 1; 102(5):1849-56.