Score Based Co-Clustering for Binary Data

R.Gowri
Department of Computer Science
Periyar University
Salem, Tamilnadu
gowri.candy@gmail.com

R. Rathipriya
Department of Computer Science
Periyar University
Salem, Tamilnadu
rathi_priyar@periyaruniversity.ac.in

Abstract **- Most of the datasets like medical datasets, expression data, network data, sensor datasets are in binary format. This article focuses on mining the block of one's or zero's (constant co-cluster) in the binary data. It represents the likelihood characteristics among the local group of elements in the data. For this purpose a score based co-clustering approach is proposed in this article. Initially this approach is attempted on the four different synthetic datasets under noisy and noiseless environments. The experimental results are compared with existing co-clustering approaches like BiMax and xMotif algorithms. The results evidence that, the proposed approach is performing well in mining the constant co-clusters in both noisy and noiseless environments.**

Keywords: Binary data, co-clustering, score, biclustering, symmetric data, SCoC

1. INTRODUCTION

In this digital world, everything is handled as digital data, where the binary data plays a vital role. The binary data are seen in different domains. In medical field as medical dataset, disease dataset, medical images, etc. In the communication field, as network routing tables, network representation data, communication dataset, etc. In bioinformatics field as expression data, protein interactions, gene interactions, disease- host interaction, etc. In security field encryption and decryption are done in the binary level data. The current article is concentrating on the co-clustering of one's and zeros's blocks in the binary dataset is concentrated. The proposed approach has the vast scope in the above mentioned fields.

The data mining is one of the ways to analyze the knowledge in the vast data. The data mining techniques for numerical datasets and categorical datasets are different. Similarly, the techniques suitable for mining the numerical datasets are not suited for binary datasets. There are different mining approaches are available for binary data. There exist many approaches especially for co-clustering the binary datasets such as BiMax, xMotif, BiBit, BiBin, etc., in the literature. The Co-clustering represents the mining of local patterns in the data(BeatrizPontes, RaúlGiráldez, Jesús S.Aguilar-Ruiz, 2015) (Victor A. Padilha, Ricardo J. G. B. Campello, 2017). In this article the score based co-clustering approach is proposed for this purpose and the BiMax and xMotif approaches are used for comparative analysis.

This section discussed about the binary data and brief introduction about the proposed approach and its scope. It is followed by the related literature study and proposed approach. The proposed approach is explored as: explanation about the proposed score measure, basic concept about the proposed score based co-clustering (SCoC) approach, Sample illustration of the proposed SCoC, synthetic dataset generation under noise and noiseless environment, experimental analysis of proposed approach, comparative analysis with existing approach, finally summarizing the proposed approach.

2. RELATED WORKS

The Binary Inclusion-Maximal Biclustering Algorithm (Bimax) (Prelić A, Bleuler S, Zimmermann P, Wille A, Bühlmann P, Gruissem W, Hennig L, Thiele L, Zitzler E., 2006) uses binary input data. it is a divide and conquer algorithm. It recursively divides the input and searches the submatrices containing ones. It uses the dicretization threshold which is mean of minimum and maximum value of the dataset. It is aminimal time consuming algorithm. Conserved Gene Expression Motifs (xMOTIFs) (Murali T, Kasif S, 2003) is a nondeterministic algorithm that finds submatrices with simultaneously conserved genes in subsets of experimental conditions in a discrete data matrix. Bit-Pattern Biclustering Algorithm (BiBit) (Rodriguez-Baena DS, Perez-Pulido AJ, Aguilar JS, 2011) which searches for maximal biclusters in binary datasets by applying the logical AND operator over all possible gene pairs. BiBit is an enumerative algorithm that works only with binary data; require the minimum

number of rows and minimum number of columns of a bicluster as input parameters. Differentially Expressed Biclusters (DeBi) (Serin A, Vingron M. Debi, 2011) is an algorithm based on a frequent itemset approach that applies a depth-first traversal on an enumeration tree to discover hidden patterns in data. It requires discrete data. It works on binary data. The main problem found is that the most of the submatrices found by this algorithm contained more columns than the desired number.

3. PROPOSED APPROACH

3.1. Score Measure

The score measure is defined for mining the co-clusters in the binary data. It is defined based on the occurrence of 1's (0's) in the input matrix (R.Gowri, R.Rathipriya, 2017). For the given data matrix (D) the score is evaluated using the equation (1).

$$\text{score}(D_{m,n}) = \frac{\sum_{i=1}^{n} \text{freq}(d_{i,J})}{n \times m} \quad (1)$$

where $D_{m,n}$ is the data matrix of order 'm x n' with binary value ('0' and '1'), the numerator represents the frequency of '1's in the given data matrix (D). The denominator represents the number of elements in the data matrix. The score depicts the density of the data matrix which is the ratio of sum of the elements to the number of elements. Here, it is same as the mean of entire matrix, as it is binary data it represents the distribution of binary value in the data matrix. The score value can be ranges from 0 to 1.

3.2. Score based Co-Clustering (SCoC)

In the previous section the score measure and its relationship with different types of co-clusters are discussed. This section discusses about the proposed Score based Co-Clustering (SCoC) approach (R.Gowri, R.Rathipriya, 2017). The SCoC is defined for the symmetric matrix and for the normal matrix separately. The Implementation I is for symmetric matrix and Implementation II is for the normal matrix. The Implementation sections discuss about the steps involved in the proposed methodology, a sample illustration of the SCoC, its experiment on the synthetic dataset and result analysis & discussion. The score based co-clustering approach adapts different threshold and testing condition for different types of co-clusters, those algorithmic changes are represented in the table 1.

Table 1. Algorithmic Setup based on Co-Cluster Types

Co-Cluster Types	**Algorithm Part**		
	Score Threshold	**Testing condition**	**Removal**
Constant Co-Cluster (1's)	1	Score< threshold	Low score
Constant Co-Cluster (0's)	0	Score>threshold	High score
Coherent Co-cluster	[0.4,0.6]	Score in range and acv < 0.9	Score not in range

3.3. SCoC for Symmetric Matrix

3.3.1. Methodology

The symmetric matrix is a square matrix where the upper right triangular matrix is same as the lower left triangular matrix. Usually the symmetric matrix is used to represent the relationship matrix of objects with one another. This implementation is especially for the symmetric matrix named as $SCoC_{sym}$. The Co-Clustering of symmetric matrix focuses only on one dimension. Here, row is concentrated. The score threshold plays vital role for mining different types of co-clusters based on the requirement. The $SCoC_{sym}$ has the following steps.

1. Evaluate the score of C
2. If score is less than threshold continue the following
    i. Evaluate row scores of C
    ii. Remove the row (and corresponding column) with low score
    iii. Evaluate score
3. Return C

The row score represent the score of each row in the matrix. The above steps are used for mining the constant 1's co-cluster from given data matrix. For the 0's co-cluster the score should be greater than threshold (0) and similarly remove the row and column with high score which is represented in the table. This approach mines the maximal size constant co-cluster from the given datamatrix.

3.3.2. Sample Illustration

The sample illustration for mining 1's constant co-cluster is given in this section. Let the Input Matrix C is as follows. The score threshold used in this sample illustration is '1'.

0 0 0 0 0 0 0 0
0 1 1 1 1 1 0 0
0 1 1 1 1 1 0 0
0 1 1 1 1 1 0 0
0 1 1 1 1 1 0 0
0 1 1 1 1 1 0 0
0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0

Table 2. Sample Illustration of the proposed approach

Iteration	Input Matrix	Row Score		Removal	Result Matrix	Score
1	0 0 0 0 0 0 0 0 0 1 1 1 1 1 0 0 0 1 1 1 1 1 0 0 0 1 1 1 1 1 0 0 0 1 1 1 1 1 0 0 0 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0.625 0.625 0.625 0.625 0.625 0 0	Min	Row & Col→1	1 1 1 1 1 0 0 1 1 1 1 1 0 0 1 1 1 1 1 0 0 1 1 1 1 1 0 0 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0.510
2	1 1 1 1 1 0 0 1 1 1 1 1 0 0 1 1 1 1 1 0 0 1 1 1 1 1 0 0 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0.714 0.714 0.714 0.714 0.714 0 0	Min	Row & Col→6	1 1 1 1 1 0 1 1 1 1 1 0 1 1 1 1 1 0 1 1 1 1 1 0 1 1 1 1 1 0 0 0 0 0 0 0	0.694
3	1 1 1 1 1 0 1 1 1 1 1 0 1 1 1 1 1 0 1 1 1 1 1 0 1 1 1 1 1 0 0 0 0 0 0 0	0.833 0.833 0.833 0.833 0.833 0	Min	Row & Col→6	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	1
4	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	Score == threshold, Stop Iterating, return C				

The sample illustration of the proposed approach is shown in the table 2. Here the 5 x 5 constant 1's co-cluster is embedded in the 8 x 8 matrix. In every iteration, the row score of the input matrix is evaluated; the row with low score and its corresponding column are removed; the score of the result matrix is evaluated. For the constant 1's co-cluster the iteration continues until the score value is 1. The score limits for the different types of co-clusters are discussed earlier. This approach is then attempted on the synthetic datasets for their performance analysis in the succeeding sections.

3.4. Dataset Description

The proposed approach is attempted on the synthetic binary datasets. In this article four different binary datasets are generated. The input data matrix is generated with implanted co-clusters and noise in it. These synthetic datasets are generated for mining constant co-clusters in the symmetric data matrix.

3.4.1. Synthetic Dataset: I

Initially, 500 x 500 input data matrix C with all the elements are zeros is generated; then a constant 1's co-cluster of size 50x50 is generated. Then implant the co-cluster in 'C' randomly and symmetrically in the data matrix C without noise as shown in the figure 1. The red portions in the figure represent the embedded bicluster.

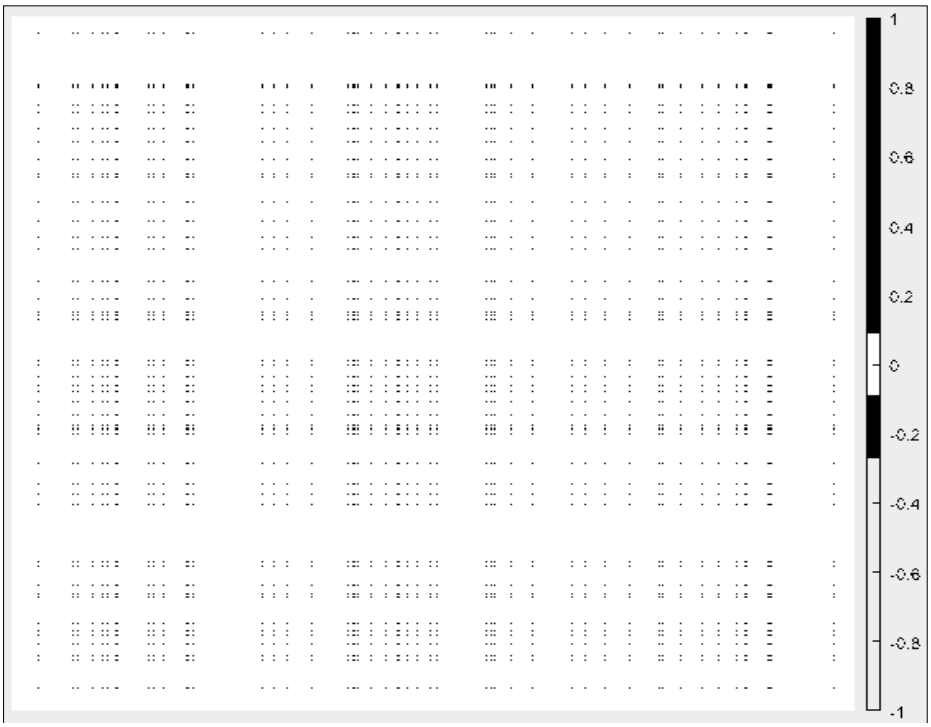
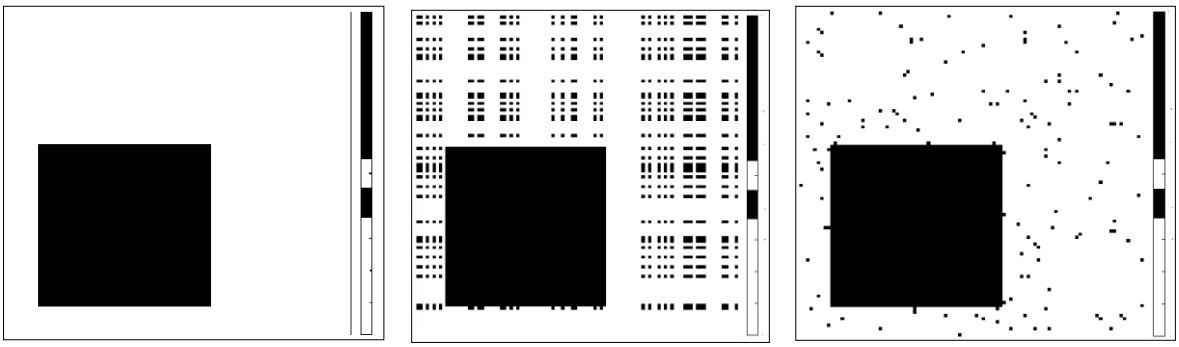


Figure 1. Synthetic Dataset I

3.4.2. Synthetic Dataset : II

Initially, 100 x 100 input data matrix C with all the elements are zeros is generated; then a constant 1's co-cluster of size 50x50 is implanted in 'C' symmetrically at specific portion in the data matrix C as shown in the figure 2. Besides, the symmetric noise and random noise is added to the dataset.

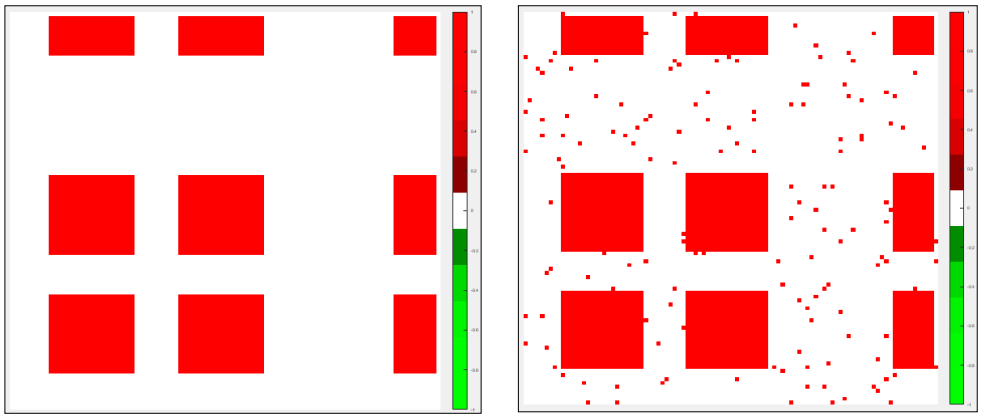


(a) (b) (c)

Figure 2. Synthetic Dataset II – (a) constant co-cluster at specific portion, (b) added symmetric noise and (c) added symmetric random noise

3.4.3. Synthetic Dataset : III

Similarly for the next dataset III, 100 x 100 input data matrix C with all the elements are zeros is generated; then a constant 1's co-cluster of size 50x50 is then implanted in 'C' symmetrically at the random portions in the data matrix C as shown in the figure 3. The random noise is added in the dataset.

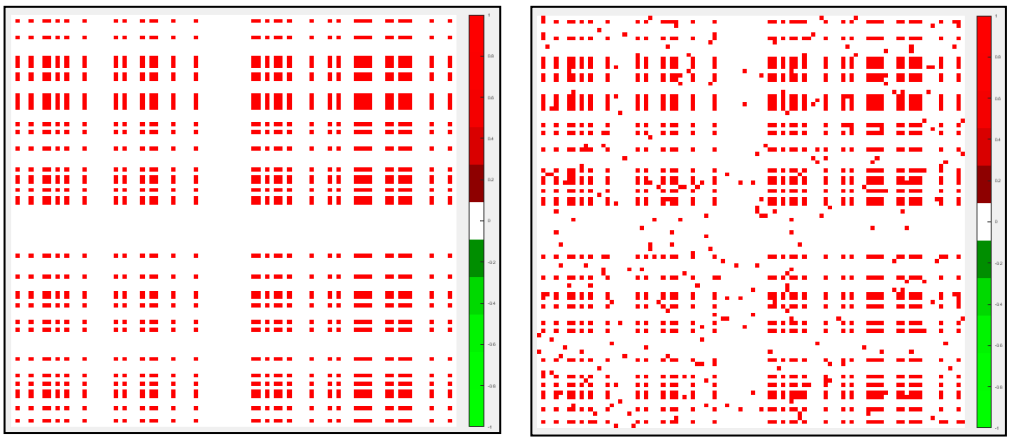


(a) (b)

Figure 3. Synthetic Dataset III – (a) constant co-cluster placed at random portions, (b) added symmetric random noise

3.4.4. Synthetic Dataset: IV

This dataset is same as dataset III, the co-cluster of size 50x50 is implanted in 'C' symmetrically and randomly in the data matrix C with random noise as shown in the figure 4.



(a) (b)

Figure 4. Synthetic Dataset IV– (a) constant co-cluster implanted randomly, (b) added symmetric random noise

3.5. Match Score Measure

The match score is used for evaluating the similarity between two biclusters (Prelić A., Bleuler S, Zimmermann P, Wille A, Bühlmann P, Gruissem W, Hennig L, Thiele L, Zitzler E, 2006) (Xiaowen Liu, Lusheng Wang 2007). This measure used in this article for comparing the implanted bicluster with the extracted bicluster. This measure exposes the performance of the different approaches discussed in this article. The match score between the biclusters is evaluated using the equation (2).

$$S(B_1, B_2) = \frac{|I_1 \cap I_2| + |J_1 \cap J_2|}{|I_1 \cup I_2| + |J_1 \cup J_2|} \quad (2)$$

where, $B_1(I_1, J_1)$ and $B_2(I_2, J_2)$ are two biclusters to be compared, I_1, I_2, J_1, J_2 represents the identifies of rows and columns in the data matrix corresponds to the biclusters. This score represents ratio of the number of common rows and columns to the number of rows and columns in the biclusters. The value of this measure varies between 0 and 1, where the value towards 1 represents the maximum similarity and towards 0 represents the dissimilarity between the biclusters.

4. RESULTS AND DISCUSSION

The proposed approach SCoCsym is attempted on four different synthetic datasets with score threshold 1 for mining constant 1's co-cluster. The performance of this SCoCsym approach is studied based on the accuracy of the outcome. Then it is compared with the performance of the existing approaches like BiMax and xMotif Biclustering algorithms for Binary data. This experiment is carried out in the Matlab environment. The biclustering algorithms are experimented using the MTBA toolbox in Matlab. The performance of these algorithms is compared based on their match score and computational time as tabulated in the table 3.

Table 3. Comparative Analysis based on Match Score Measure

	Syn_Data1	**Syn_Data2**		**Syn_Data3**		**Syn_Data4**	
	Without Noise	**Without Noise**	**With Noise**	**Without Noise**	**With Noise**	**Without Noise**	**With Noise**
SCoC	1	1	1	1	1	1	1
BiMax	1	1	0.4857	1	0.5000	1	0.4810
xMotif	0.0460	0.9174	0.0848	0.25	0.4552	0.1850	0.0680

The match between the implanted co-cluster and extracted co-cluster based on their row and column identifiers is evaluated. The match score value lies between 1 and 0, where 1 represents the highest match and the value towards 0 represents fewer matches between the co-clusters. The match score of the results shows that the proposed approach $SCoC_{sym}$ is outperforming the existing co-clustering approaches for binary data. The implanted co-clusters are extracted even under the noisy space. The randomly implanted co-clusters are also extracted successfully by the proposed approach. The BiMax algorithm also extracts the co-clusters only in the dataset without noise. The presence of the noise affects the performance of the existing approaches. This table summarizes the overall performance of the proposed approach.

5. CONCLUSION

This article focused on mining the block of one's or zero's (constant co-cluster) in the binary data. The score based co-clustering (SCoC) approach is proposed in this article. Initially this approach is attempted on the four different synthetic datasets under noisy and noiseless environments. The existing biclustering approaches like BiMax and xMotif algorithms are choosen for comparative study. The experimental results evidence that, the proposed approach is performing well in mining the constant co-clusters in both noisy and noiseless environments than the existing approach. This research work can be further extended for mining the coherent co-clusters; for mining co-clusters in rectangular matrix; can be attempted on real-time datasets for its betterment. This approach can be applied for mining differently expressed genes; for mining mass in the image analysis; for mining more connected regions in the networks and so on.

6. ACKNOWLEDGMENT

The first author acknowledges the UGC for the financial support to her research under the UGC NET JRF (Student Id: 3384/(OBC)(NET JULY-2016)) Scheme.

REFERENCES

BeatrizPontes, RaúlGiráldez, Jesús S.Aguilar-Ruiz (2015), Biclustering on expression data: A review, Journal of Biomedical Informatics, 57, 163-180

Rodriguez-Baena DS, Perez-Pulido AJ, Aguilar JS (2011). A biclustering algorithm for extracting bit-patterns from binary datasets. Bioinformatics, 27(19), 2738–45.

Murali T, Kasif S (2003). Extracting conserved gene expression motifs from gene expression data. In: Pacific Symposium on Biocomputing. Stanford: Stanford Medical Informatics: 2003, 77–88.

Serin A, Vingron M (2011). Debi: Discovering differentially expressed biclusters using a frequent itemset approach. Algorithms Mol Biol. 6(1), 18

Prelić A, Bleuler S, Zimmermann P, Wille A, Bühlmann P, Gruissem W, Hennig L, Thiele L, Zitzler E (2006). A systematic comparison and evaluation of biclustering methods for gene expression data. Bioinformatics. 22(9), 1122–9.

R.Gowri and R.Rathipriya (2017), Cohesive Sub-Network Mining in Protein Interaction Networks using Score based Co-Clustering with MapReduce Model (MR-CoC), ICACA 2017, Karunya University.

Xiaowen Liu, Lusheng Wang (2007), Computing the maximum similarity bi-clusters of gene expression data, *Bioinformatics*, 23(1), 50–56

Victor A. Padilha, Ricardo J. G. B. Campello (2017), A systematic comparative evaluation of biclustering techniques, BMC Bioinformatics.