Biclustering of Time Series Data using Factor Graphs

Matteo Denitto
University of Verona
Department of Computer Science
Verona, Italia
matteo.denitto@univr.it

Alessandro Farinelli
University of Verona
Department of Computer Science
Verona, Italia

Manuele Bicego
University of Verona
Department of Computer Science
Verona, Italia

ABSTRACT

Biclustering regards the simultaneous clustering of both rows and columns of a given data matrix. A specific application scenario for biclustering techniques concerns the analysis of gene expression time-series data, wherein columns dataset are temporally related. In this context, biclustering solutions should involve subset of genes sharing 'similar' behaviours among *consecutive* experimental conditions. Due to the intrinsic spatial constraint required by time-series dataset, current Factor Graph (FG) based approaches cannot be applied. In this paper we introduce *Time-Series constraints* forcing biclustering solution to have contiguous columns. We optimize the model by using the Max-Sum algorithm, whose message update rules have been derived exploiting The Higher Order Potentials (THOP). The proposed method has been assessed on a real world dataset and the retrieved biclusters show that it can provide accurate and biologically relevant solutions.

CCS Concepts

•**Applied computing** → **Bioinformatics;**

Keywords

Biclustering, Time Series, Factor Graph, Max-Sum, Gene Expression,

1. INTRODUCTION

Biclustering, widely known also as co-clustering, refers to the simultaneous clustering of both rows and columns of a given data matrix [7]. The goal of biclustering techniques is to retrieve groups of rows which share 'similar' behaviours in a subset of columns, and viceversa. Compared to clustering, biclustering focuses on local correlations between rows/columns portions instead of considering the whole trend. Its origins and the main application field is Bioinformatic, namely in the analysis of gene expression data [7]. A particular task in gene expression scenario is represented by the analysis of time series data. In these datasets, samples refer to experimental conditions which are temporally related (e.g. growth stages of a plant, disease evolution). Biclustering approaches in this scenario focus on the identification of genes that act similarly in consecutive subset of samples [8]. This introduces a spatial constraint that cannot be satisfied by classical biclustering approaches, Figure 1 provides an intuitive example.

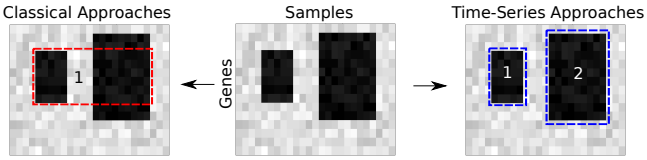

Figure 1: Biclustering of Time Series Data. The picture shows how the data matrix (on top) would be analysed by classical biclustering approaches (left), and how dedicated Time-Series biclustering approaches retrieve the desired information.

Recently, a trend of research focuses on the investigation of *Factor Graphs* (FG) to face the biclustering problem [4]. The most recent FG based algorithm [4] redefines the biclustering problem as the iterative search of one bicluster, and it provides a model that define a mask on the data matrix highlighting which points belong to the solution and which not. Such mask is obtained by optimizing the objective function described in the FG with the *Max-Sum algorithm*. Although the method proposed in [4] improves the scalability of FG based approaches, it cannot be applied on time-series dataset. One of the advantages in using FG approaches is the fact that these models can be easily modified by plugging in new factors, given that corresponding Max-Sum messages can be derived efficiently. In this paper we extend the model proposed in [4] by introducing a constraint enforcing the solution columns to be contiguous. We derive the new messages update rules employing The Higher Order Potentials [9], and we plug them in the pipeline described in [4]. The proposed technique does not alter the time and space complexity of [4], hence preserving the scalability improvement reached by such method. We conduct a simple test on a real gene expression matrix containing time-series data: preliminary results show that the investigation of FGs for biclustering time-series data provides accurate and biologically significant solutions.

FGs and the Max-Sum Algorithm.

A FG is a bipartite graph graph including: i) *variable nodes*,

SAC 2017,April 03-07, 2017, Marrakech, Morocco

http://dx.doi.org/10.1145/3019612.3019848

one for problem variable, usually drawn as circles and ii) *factor nodes*, one for each local function, usually drawn as squares. Such graph is bipartite in the sense that there are only edges/connections between factors and variable nodes (not between two variables or two factors). The obtained graph represents how the global objective function decomposes in smaller local functions (factors) providing a model which is easier to analyse. There are no limitations on the form of the local functions, in fact factors can be designed to impose hard or soft constraints, and they may involve different types of variables (*e.g.*, categorical, integer, binary, real numbers). However, depending on the optimization algorithm adopted, the choices made when designing a FG have a drastic effect on the difficulty of the resulting optimization problem [4]. In this paper, due to its recent exploitation in the biclustering field, we adopt the widely known *Max-Sum* algorithm [1] [4].

2. BICLUSTERING FG FOR TIME SERIES

Components, Variables and Factors.

1. *Biclustering can be seen as the incremental search for the largest bicluster.* As in [4] the solution of our FG is a binary 'mask' C having the same size of the matrix under analysis, each variable $c_{ij} \in \{0, 1\}$ (with $1 \leq i \leq \#Rows$ and $1 \leq j \leq \#Columns$). Each variable c_{ij} is equal to one if, and only if, an entry a_{ij} belongs to the solution.

2. *Entries value counts.* Biclusters with high valued entries should be preferred since they are related with genes over-expressions. To represent this aspect in the objective function the model comprehend one factor $A_{ij}(c_{ij})$ for each variable, which sums the value a_{ij} to the objective if, and only if, the variable c_{ij} has been set to 1.

3. *Biclusters are coherent.* Coherence is the crucial aspect that distinguish a bicluster from a random submatrix. In order to obtain the desired solutions, we penalize the incoherent ones introducing a factor on each pair of variables. Such factor, called $O_{ijtk}(c_{ij}, c_{tk})$, subtracts from the objective function the amount of incoherence between two variables $I(a_{ij}, a_{tk})$ if both of them have been set to 1. In this paper we adopt as coherence criteria the simple difference among all the entries belonging to a bicluster: $I(a_{ij}, a_{tk}) = |a_{ij} - a_{tk}|$.

4. *Biclusters are sub-matrices.* This means that once we select the entries belonging to the solution, we must be sure that these represent a valid assignment (*i.e., form a sub-matrix*). For this reason the FG involves a set of factors defined over each couple of rows (or columns) $B_i t(c_{i1}, \ldots, c_{im}, c_{t1}, \ldots, c_{tm})$, that enforce the selected entries belonging to have a rectangular shape (i.e., form a sub-matrix).

5. *Time series solution must involve contiguous experimental conditions.* As mentioned in Section 1, the analysis of time-series data focus on how genes patterns evolve in consecutive instant of times. Returning to Figure 1, classical biclustering approaches cannot retrieve such information. Hence we introduce a set of spatial constraint in order to devise a suitable one. Given a set of variable binary variables x, we define the *Time-Series* constraint $TS(x)$ such that the ones in x do not have zeros in between or, more formally, must form a contiguous subset. Assuming that the columns of the data matrix represent the experimental conditions, if we apply such constraint on each row of C the solution is enforced to have contiguous columns and hence to contain a time-series bicluster.

Concerning messages, in this short paper we report only the equations adopted to encode the time-series constraint in the model. For a detailed explanation about messages derivation please refer to [5]. Specifically: messages going from variables to functions are can be computed as:

$$\gamma_{ij} = \sum_{k=1}^{m} \eta_{ij}^{k} + \sum_{tk=(1,1)}^{(n,m)} \sigma_{ij}^{tk} + \alpha_{ij}$$

where $\eta_{ij}^{k}, \sigma_{ij}^{tk}$ and α_{ij} are incoming messages for the variable c_{ij}. Regarding the message going from a factor node TS_i to a variable node c_{ij}, we derive it exploiting a particular THOP called *convex-set potential* [9]; and it can be computed as:

$$\begin{aligned}\rho_{ij} =& MS(c_{1:j-1}, c_{j-1} = 1) + MS(c_{j+1:N}, c_{j+1} = 1) + \\ & - \max\left(MS(c_{1:j-1}), MS(c_{j+1:N})\right).\end{aligned}$$

where $MS(c_{1:j-1}, c_{j-1} = 1)$ is a function retrieving the maximum weighted contiguous subsequence in the subset $\{c_1, \ldots, c_{j-1}\}$ and forcing c_{j-1} to be equal to 1 (if the second part is missing, no variables are constrained).

3. EXPERIMENTAL EVALUATION

The algorithm performances have been assessed on a Real Gene Expression dataset, namely the CHO *et al* yeast cell-cycle dataset [3]. The dataset is composed by a matrix where 6457 genes have been sampled in 17 consecutive time steps with an interval of 10 minutes.

Although recent models improved the scalability of FG based approaches, real data matrices are still far to be directly analysed [4]. In our case the bottleneck is represented by the space complexity required by the model (*i.e.,* the coherence is calculated for each pair of entries with complexity $\mathcal{O}(n^2m^2)$). Hence, as proposed in [4], we run the algorithm on randomly extracted sub-matrices where about 120 rows have been selected. First, to reduce the matrix dimensionality we applied a variance-based gene selection, as already adopted in relevant literature [2]. Each row of the obtained matrix have been then rescaled such that the 2-sigma interval lies in $[0, 1]$, and missing values have been recovered using the method proposed in [10]. We retrieve one bicluster from each sub-matrix and then the following evaluation criteria have been adopted:

1. *Mean Square Residue (MSR)* - it assess the fluctuation of expression level for all rows in the bicluster. The

[1]Max-Sum is a message passing technique based on the definition of two functions, called *messages*, exchanging information between connected nodes in the graph. These messages are iteratively exchanged until a convergence criterion is met (commonly defined over the variables configuration or the objective function). For a detailed perspective on the Max-Sum algorithm including pros and cons please refer to [5].

ID	#Rows	#Cols (first-last)	MSR	GO-Terms
1	20	3 (14 - 16)	0.0017	GO:0065008 GO:0022890
2	18	4 (12 - 15)	0.0011	GO:0042254 GO:0022613
3	18	3 (14 - 16)	0.0007	GO:0009063 GO:0044270
4	17	4 (6 - 9)	0.0015	GO:0005515 GO:0051649
5	17	4 (6 - 9)	0.0015	GO:0008202 GO:0016125
6	17	3 (11 - 13)	0.0008	GO:0065008 GO:0022890
7	17	3 (6 - 8)	0.0013	GO:0031974 GO:0043233
8	16	3 (3 - 5)	0.0005	GO:0031980 GO:0005759
9	16	3 (7 - 9)	0.0016	GO:0016020 GO:0031090
10	16	3 (4 - 6)	0.0017	GO:0005730 GO:0015078

Table 1: Top 10 largest Biclusters obtained by the proposed approach on the yeast dataset.

smaller the MSR, the higher the correlation. Given a bicluster A_{TK}, the MSR is computed as

$$MSR(TK) = \frac{1}{|T||K|} \sum_{t \in T} \sum_{k \in K} (a_{tk} - a_{Tk} - a_{tK} + a_{TK})^2$$

where a_{Tk} is the mean of the k^th column in the bicluster, a_{tK} is the mean of the t^th row in the bicluster, and a_{TK} is the mean of the whole bicluster.

2. *Gene Ontology (GO) terms* - are a fundamental qualitative information that highlights whether the genes contained in a certain bicluster are biologically related or not. The GO enrichment analysis have been performed via the *GOstat* online application[2] [1] setting as p-value threshold 0.05.

Table 1 reports the results concerning the top 10 largest biclusters, it shows for each bicluster: the number of rows composing the bicluster, the columns interval selected, the MSR and the top 2 GO terms according to the p-value and number of genes involved. It can be seen by the MSR and GO term columns that the retrieved biclusters present high coherence and they are significantly and biologically relevant.

Surely it would be interesting to compare the proposed approach with other state-of-the-art techniques such as [8], however with these preliminary results we want to demonstrate that the extension of FG based approaches to time-series dataset is sound and, moreover, it provides accurate and biologically meaningful solutions. To make results comparable with current state-of-the-art a first step could be trying to merge the results increasing the biclusters size. A possible solution can be the following: i) given the set of bicluster previously obtained, merge all the rows of the biclusters sharing the same column subset; ii) then evaluate the FG objective function for each novel bicluster and, if the value is bigger then the sum of the previous ones, replace them. Table 2 presents two bicluster obtained with the heuristic just described. The MSR of the retrieved biclusters is comparable with the performances presented in Table 1, demonstrating that the proposed approach would provide accurate solution also in bigger data matrices. Moreover grouping the results allows to retrieve different and more informative GO terms, specifically in the GO term highlighted in bold 31 of the 38 selected genes are involved in the same biological process.

[2]http://gostat.wehi.edu.au/

ID	#Rows	#Cols (first-last)	MSR	GO-Terms
11	38	3 (14 - 16)	0.0013	**GO:0005737** GO:0022890
12	57	4 (6 - 9)	0.0015	GO:0044425 GO:0005515

Table 2: Result of merged biclusters. Highlighted in bold a GO term that was not enriched in basic results, this GO term involve 31 of the 38 genes in the bicluster.

4. CONCLUSIONS

In this paper we propose a FG based approach dedicated to time-series dataset. The algorithm presented is an extension where the model proposed in [4] has been modified introducing Time-Series constraints enforcing the solution to have contiguous columns. We implement Max-Sum messages update rules exploiting The Higher Order Potentials, the resulting messages do not alter the complexity of the previous model. We test the approach on a real gene expression dataset and the biclusters obtained present high coherence and are biologically consistent. Future works will surely regard the comparison with other state-of-the-art time-series algorithms.

5. REFERENCES

[1] T. Beissbarth and T. P. Speed. GOstat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics*, 2004.

[2] M. Bicego, P. Lovato, A. Perina, M. Fasoli, M. Delledonne, M. Pezzotti, A. Polverari, and V. Murino. Investigating topic models' capabilities in expression microarray data classification. *TCBB*, 2012.

[3] R. J. Cho, M. J. Campbell, E. A. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T. G. Wolfsberg, A. E. Gabrielian, D. Landsman, D. J. Lockhart, et al. A genome-wide transcriptional analysis of the mitotic cell cycle. *Molecular cell*, 1998.

[4] M. Denitto, A. Farinelli, and M. Bicego. Biclustering gene expressions using factor graphs and the max-sum algorithm. In *IJCAI*, 2015.

[5] M. Denitto, A. Farinelli, M. Figueiredo, and M. Bicego. A biclustering approach based on factor graphs and the max-sum algorithm. *PR*,2017.

[6] B. Frey and D. Dueck. Clustering by passing messages between data points. *Science*, 2007.

[7] S. Madeira and A. Oliveira. Biclustering algorithms for biological data analysis: a survey. *IEEE TCBB*, 2004.

[8] S. C. Madeira, M. C. Teixeira, I. Sa-Correia, and A. L. Oliveira. Identification of regulatory modules in time series gene expression data using a linear time biclustering algorithm. *IEEE/ACM TCBB*, 2010.

[9] D. Tarlow, I. E. Givoni, and R. S. Zemel. Hop-map: Efficient message passing with high order potentials. In *AISTATS*, 2010.

[10] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman. Missing value estimation methods for dna microarrays. *Bioinformatics*, 2001.