HAL
open science

Formal Concept Analysis for Identifying Biclusters with Coherent Sign Changes

Nyoman Juniarta, Miguel Couceiro, Amedeo Napoli

To cite this version:

Nyoman Juniarta, Miguel Couceiro, Amedeo Napoli. Formal Concept Analysis for Identifying Biclusters with Coherent Sign Changes. 2019. hal-02181600

HAL Id: hal-02181600

https://inria.hal.science/hal-02181600v1

Preprint submitted on 12 Jul 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Formal Concept Analysis for Identifying Biclusters with Coherent Sign Changes

Nyoman Juniarta*, Miguel Couceiro*, Amedeo Napoli*

*Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
nyoman.juniarta@loria.fr

Abstract. In this paper we are studying the task of finding coherent-sign-changes biclusters in a binary matrix. This task can be applied to the interpretation of gene expression data, where such a bicluster represents a set of experiments that affect a set of genes in a consistent way. We start with a binary table and study biclustering methods based on FCA and partition pattern structures. Pattern concepts provide biclusters and their hierarchical relation, which can be used to analyze the profile of genes in the given expression data. Our approach is purely symbolic, so we can detect larger biclusters and work with rather complex data.

1 Introduction

Gene expression data can be represented as a matrix, where rows and columns represent genes and experiments respectively. Each cell contains the numeric expression level of a given gene under a given experiment. In such data, we can say that an experiment affect a gene by either lowering or raising its expression, according to the gene's normal level. One may be interested in finding a subset of genes and a subset of experiments, such that the experiments affect the genes in a consistent way. In other words, any two experiments in the subset have always either the same effect or the opposite effect on every gene in the subset. This task corresponds to the mining of coherent-sign-changes (`CSC`) biclusters.

Biclustering is an important technique aimed at discovering patterns in a matrix representing a dataset. It is related to standard clustering whose main objective is to group the rows based on their similarity. On the other hand, biclustering refers to the problem of discovering submatrices whose cells exhibit similar behavior. This problem is also called co-clustering Govaert and Nadif (2013), where rows and columns are clustered simultaneously.

In this paper, we present a method based on FCA and pattern structures for discovering a specific type of bicluster: coherent-sign-changes bicluster. An existing approach in Tanay et al. (2002) can mine this bicluster type, but it is statistical, since its discovery of `CSC` biclusters is based on the magnitude of the expression changes. Our approach is more symbolic, by taking into account only the direction of the changes, with expectation of detecting larger biclusters. Our FCA-based method also gives us the hierarchical structure of all biclusters, allowing an easier interpretation of the results by experts. Furthermore, pattern structures and AddIntent algorithm allows us to define a threshold of bicluster size, so that we can limit the amount of retrieved biclusters.

2	2	2	2		+	+	−
2	2	2	2		+	+	−
2	2	2	2		−	−	+
2	2	2	2		−	−	+
	(a)					(b)	

TAB. 1 – *Examples of two bicluster types: (a) constant-values and (b) coherent-sign-changes (CSC).*

2 Related Work

The row–column clustering was introduced in Hartigan (1972), and Cheng and Church Cheng and Church (2000) were the firsts to used the term biclustering while working on gene expression data. A bicluster in Cheng and Church (2000) is a subset of genes and a subset of conditions with a high similarity score, statistically measured by calculating variances over all values in the submatrix.

Still in the domain of gene expression data, the algorithm called SAMBA was proposed in Tanay et al. (2002) to discover a submatrix where the expressions of a subset of genes significantly changes across a subset of conditions. The first model of SAMBA searches a submatrix where there is a *joint* change across all genes, without looking whether it is an increase or a decrease. The second model takes into account the direction of the change, such that any two conditions in the submatrix have either always the same effect or always the opposite effect. We call this type of submatrix a coherent-sign-changes bicluster, as denoted in Madeira and Oliveira (2004).

Regarding bicluster discovery based on FCA, several methods were proposed. In a binary matrix, dense approximate bicluster discovery was studied in Gnatyshak et al. (2012); Ignatov et al. (2012a) based on standard FCA. This is similar to mining formal concept, but instead of "exact" concepts, the authors relax the problem such that the "approximate" concepts (having a certain amount of empty cells) can also be detected. For biclustering with similar values in a numerical matrix, Kaytoue et al. in Kaytoue et al. (2011) proposed standard FCA with scaling and interval pattern structures. Triadic Concept Analysis was also studied in Kaytoue et al. (2014) to extract this bicluster type. Furthermore, a partition pattern structure was presented in Codocedo and Napoli (2014); Kaytoue (2011) for mining bicluster with constant columns.

3 Biclustering

We consider that a dataset is composed of a set of objects G, each of which has values over a set of attributes M. This dataset can be represented as a numerical context (G, M, I) where G is a set of objects, M is a set of attributes, and I corresponds to $m(g)$, which is the value of $m \in M$ for object $g \in G$.

One may be interested in finding which subset of objects possesses the same values w.r.t. a subset of attributes. Regarding the matrix representation, this is equivalent to the problem of finding a submatrix that has a constant value over all of its elements (example in Table 1a). This task is called biclustering with constant values, which is a simultaneous clustering of the rows and columns of a matrix.

TAB. 2 – *Example.*

$\mathcal{M}$	m_1	m_2	m_3	m_4
g_1	$+$	$+$	$-$	$-$
g_2	$+$	$+$	$-$	$-$
g_3	$-$	$-$	$+$	$-$
g_4	$+$	$+$	$+$	$+$
g_5	$-$	$-$	$-$	$-$

In coherent-sign-changes (CSC) bicluster, the matrix is binary. In this bicluster, each row is correlated (either entirely identical or entirely opposite) to all other rows. In the example in Table 1c, the first row is identical to the second and opposite to the third and fourth. We can also see this bicluster by comparing the columns. In the example, the first column is identical to the second and opposite to the third.

In a binary dataset (G, M, I), given a set of objects $A \subseteq G$ and an attribute $m \in M$, $m(A)$ is the *column submatrix* formed by the attribute m over A. The submatrix $m_j(A)$ is equal to $m_k(A)$, denoted as $m_j(A) \simeq m_k(A)$, if all rows in $m_j(A)$ are either entirely identical or entirely opposite to the corresponding rows in $m_k(A)$. With the previous notation, given a binary dataset (G, M, I), a pair (A, B) (where $A \subseteq G$, $B \subseteq M$) is a *coherent-sign-changes bicluster* if $\forall m_j, m_k \in B : m_j(A) \simeq m_k(A)$.

4 The Pattern Structures of Signed Partition

In the task of CSC bicluster discovery in a formal context (G, M, I), we propose an approach based on partition pattern structures. Instead of partition of objects in G as described in Baixeries et al. (2014); Codocedo and Napoli (2014), here we use *partition of attributes* in M. It is still similar to an object partition since an attribute partition covers every attribute in M and there is no overlapping between any two partition components.

In the original partition pattern structures, a *partition* is a set of components, where a *component* is a set of elements. For example, $\{\{a, b, c\}, \{d, e\}\}$ is a partition with two components. The first component has three elements, while the second has two. For CSC biclustering, we define a component as a set of signed elements. This gives us $\{\{a^+, b^-, c^+\}, \{d^-, e^+\}\}$ as an example of our component. In this work, we define the similarity operator among our components. Based on this similarity, we can obtain a *signed partition pattern concept*, which eventually gives us CSC biclusters.

The number of concepts can be exponential. Our experiments show that these set of concepts and the computational time can be reduced by introducing parameters. For example, we can choose a threshold that defines a minimum number of columns or rows of the biclusters.

References

Baixeries, J., M. Kaytoue, and A. Napoli (2014). Characterizing functional dependencies in formal concept analysis with pattern structures. *Annals of Mathematics and Artificial Intelligence 72*, 129–149.

Cheng, Y. and G. M. Church (2000). Biclustering of expression data. In *ISMB*, Volume 8, pp. 93–103.

Codocedo, V. and A. Napoli (2014). Lattice-based biclustering using partition pattern structures. In *Proceedings of the Twenty-first European Conference on Artificial Intelligence*, pp. 213–218. IOS Press.

Codocedo-Henríquez, V. (2015). *Contributions à l'indexation et à la récupération d'information utilisant l'analyse formelle de concepts*. Ph. D. thesis, Université de Lorraine.

Gnatyshak, D., D. I. Ignatov, A. Semenov, and J. Poelmans (2012). Analysing online social network data with biclustering and triclustering. In *Proceedings of the "Concept Discovery in Unstructured Data" conference*, Volume 871, pp. 30–39. Citeseer.

Govaert, G. and M. Nadif (2013). *Co-clustering*. Wiley-IEEE Press.

Hartigan, J. A. (1972). Direct clustering of a data matrix. *Journal of the american statistical association 67*(337), 123–129.

Ignatov, D. I., S. O. Kuznetsov, and J. Poelmans (2012a). Concept-based biclustering for internet advertisement. In *Data Mining Workshops (ICDMW), 2012 IEEE 12th International Conference on*, pp. 123–130. IEEE.

Ignatov, D. I., J. Poelmans, and V. Zaharchuk (2012b). Recommender system based on algorithm of bicluster analysis RecBi. *arXiv preprint arXiv:1202.2892*.

Kaytoue, M. (2011). *Traitement de données numériques pas analyse formelle de concepts et structures de patrons*. Ph. D. thesis, Université Henri Poincare – Nancy 1.

Kaytoue, M., S. O. Kuznetsov, J. Macko, and A. Napoli (2014). Biclustering meets triadic concept analysis. *Annals of Mathematics and Artificial Intelligence 70*(1-2), 55–79.

Kaytoue, M., S. O. Kuznetsov, A. Napoli, and S. Duplessis (2011). Mining Gene Expression Data with Pattern Structures in Formal Concept Analysis. *Information Science 181*(10), 1989–2001.

Madeira, S. C. and A. L. Oliveira (2004). Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB) 1*(1), 24–45.

Tanay, A., R. Sharan, and R. Shamir (2002). Discovering statistically significant biclusters in gene expression data. *Bioinformatics 18*(suppl_1), S136–S144.