Scalable Biclustering Algorithm Considers the Presence or Absence of Properties

Abdelilah Balamane, Statistic Canada, Canada

ABSTRACT

Most existing biclustering algorithms take into account the properties that hold for a set of objects. However, it could be beneficial in several application domains such as organized crimes, genetics, or digital marketing to identify homogeneous groups of similar objects in terms of both the presence and the absence of attributes. In this paper, the author proposes a scalable and efficient algorithm of biclustering that exploits a binary matrix to produce at least three types of biclusters where the cell's column (1) are filled with 1's, (2) are filled with 0's, and (3) some columns filled with 1's and/or with 0's. This procedure is scalable and it's executed without having to consider the complementary of the initial binary context. The implementation and validation of the method on data sets illustrates its potential in the discovery of relevant patterns.

KEYWORDS

Absence, Attributes, Biclustering, Binary, Concept, Formal, Negative, Patterns, Scalable

1. INTRODUCTION

Gene expression is the process by which the instructions in our DNA are converted into a functional product, such as a protein. It is the most important source of biological data used to reveal the interaction and functionality of genes.

Biclustering is a popular technique to study and analyze gene expression data. One of its objectives is to cluster genes based on their expression under multiple conditions, or to cluster conditions based on the expression of multiple genes. This technique was introduced by Hartigan (Hartigan, 1972) who proposed a new method of matrix partitioning called biclustering or coclustering. It has allowed to simultaneously group objects (matrix's rows) and properties (matrix's columns) to create sub-matrices called biclusters where objects are highly correlated with properties. The main advantages of biclustering are the direct interpretation of sub-matrices, and the ability to identify correlations between sets of objects and sets of attributes, mainly in a large and scattered matrix. It should be noted that the high level of cohesion within the different biclusters is explained by the fact that only the relevant properties for the creation of a set of objects are used, rather than all the available properties.

However, most existing biclustering algorithms can only reveal positive gene interactions. Recent research (Arvas, et al., 2011; Ayadi, W., & Hao, J. K., 2014; Chung, et al., 2018; Nepomuceno, J. A., Troncoso, A., & Aguilar-Ruiz, J. S., 2011; Spainhour, J. C., Lim, H. S., Yi, S. V., & Qiu, P.,

DOI: 10.4018/IJDWM.2021010103

Table 1. Discretized gene expression matrix K1

K1	C1	C2	C3	C4	C5
g1	1	1	1	1	1
g2	1	1	1	1	1
g3	-1	1	-1	-1	1
g4	-1	1	-1	-1	1
g5	-1	1	-1	-1	1
g6	1	-1	-1	-1	-1

2019) shows that groups of biologically significant genes may exhibit negative correlations. The few biclustering algorithms that take into consideration such correlations employ a discretized matrix of gene expression and its complement to calculate biclusters. The temporal complexity of these algorithms is high due to the use of matrix complement. The originality of this work lies in the proposal of a generic, efficient and scalable method called BiP (Biclustering Procedure) that calculates from a binary matrix a set of biclusters showing said correlation without recourse to the complement of the matrix. Diverse types of biclusters can then be obtained with cell's content filled: (i) only with 1 (type 1), (ii) only with 0 (type 2), and (iii) columns filled with 1 and/or 0 (type 3).

The remaining sections of this paper are organized as follows: Section II illustrates an example of discretized genes expression matrix K1 from which three biclusters of type 2, 1 and 3 are extracted. Section III is dedicated to the related work. In section IV, we present an overview of the formal concept analysis and the Patricia Tree. Section V describes the algorithm BiP. In section VI we evaluate our algorithm on real and synthetic data. Finally, we conclude this paper and list future work in Section VII.

2. EXAMPLE OF BICLUSTERS PRODUCED BY BIP

Several datamining and machine learning algorithms have been proposed to unveil the interactions between genes and conditions from a gene expression matrix. These algorithms require in many cases that the interaction matrix be discretized. Our algorithm is one of them. In this section we start using Table 1 as example of discretized matrix obtained from a gene expression matrix using an appropriate method of discretization (Cristian et al., 2016). The author applied the algorithm BiP on Table 1 to produce three biclusters of type 2, 1 and 3 respectively represented in Table 2, Table 3 and Table 4.

The first bicluster of type 2 contains four genes g3g4g5g6 negatively correlated with the conditions C3C4 when the bicluster of type 1 contains the genes g1g2 positively correlated with the set of conditions C1C2C3C4C5. Finally, in the type 3 bicluster the set of genes g3g4g5 are correlated positively with C2C5 conditions and negatively with C1C3C4.

Table 2. Bicluster of type 2

Type 2	C3	C4
g3	-1	-1
g4	-1	-1
g5	-1	-1
g6	-1	-1