Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 112 (2017) 278–287

Procedia Computer Science

www.elsevier.com/locate/procedia

21th International Conference on Knowledge Based and Intelligent Information and Engineering Systems, KES2017

Mining Negative Correlation Biclusters from Gene Expression Data using Generic Association Rules

Amina Houari[a,*], Wassim Ayadi[b], Sadok Ben Yahia[a]

[a] *University of Tunis El Manar, Faculty of Sciences of Tunis, LIPAH-LR11ES14, 2092 Tunis, Tunisia*
[b] *University of Tunis, National Higher Engineering School of Tunis, LaTICE-LR11ES04, 1008 Tunis, Tunisia.*

Abstract

A majority of existing biclustering algorithms for microarrays data focus only on extracting biclusters with positive correlations of genes. Nevertheless, biological studies show that a group of biologically significant genes may exhibit negative correlations. In this paper, we propose a new biclustering algorithm, called **NBic-ARM** (**N**egative **Bic**lusters using **A**ssociation **R**ule **M**ining). Based on Generic Association Rules, our algorithm identifies negatively-correlated genes. To assess NBic-ARM's performance, we carried out exhaustive experiments on three real-life datasets. Our results prove NBic-ARM's ability to identify statistically and biologically significant biclusters.

© 2017 The Authors. Published by Elsevier B.V.
Peer-review under responsibility of KES International

Keywords: Biclustering; Negative correlations; Generic Association Rules ; Data mining; Bioinformatic; DNA microarray data.

1. Introduction

Molecular Biology research can only grow through the development of its relevant technologies. This is because, conducting research on huge numbers of genes using traditional methods has not been possible. Such research includes DNA microarray technologies. These latter help to identify new genes, discover their functions and annotations, under diverse experimental conditions. Since its invention, this technology has been applied to several biological and medical researches. To do so, gene expression data is arranged in a data matrix, where rows represent genes, columns represent samples (experimental conditions), and each cell of the matrix denotes the expression level of a gene under a certain experimental condition. To analyse this huge data, data mining techniques, especially clustering, are used. Clustering methods can be applied to either rows or columns of the gene expression data matrix, separately. However, the use of clustering algorithms has two major drawbacks: (1) They consider the whole set of samples. This is despite the fact that genes may not be relevant to every sample. Instead, they can be relevant to only a subset of samples, which is a fundamental aspect for numerous problems in the Biomedicine field[1]. Thus, clustering should be performed

[*] Corresponding author. Tel.: +216 52 695 365.
E-mail address: amina.houari@fst.utm.tn

1877-0509 © 2017 The Authors. Published by Elsevier B.V.
Peer-review under responsibility of KES International
10.1016/j.procs.2017.08.262

simultaneously on both genes and conditions. (2) Each gene can only be clustered into one group. Nevertheless, many genes can belong to several clusters depending on their influence in different biological processes[2].

In this respect, biclustering, which is a particular clustering type, came to palliate these drawbacks. Hence, biclustering aims to identify maximal sub-matrices (*aka biclusters*) where a subset of genes express highly correlated behaviors over a range of conditions[3]. Nevertheless, the biclustering task is a highly combinatorial problem and is known to be NP-Hard one[4].

According to[5], existing biclustering algorithms can be grouped into two main classes: Systematic search algorithms and stochastic search ones. Recent attempts to perform biclustering use pattern-mining searches[6,7,8]. Pattern-mining-based biclustering approaches aim to perform efficient and flexible searches with better solutions in term of coherency and quality[9]. These advantages brought these algorithms into the spotlight when it comes to biological data analysis[6,7,10,11,12,13].

Among these pattern-mining-based algorithms we can find ARM-based biclustering algorithms[14]. However, using ARM can lead to the obtainment of a huge number of redundant association rules which can be drawn even from small amount of data. To solve this problem, the Informative Generic Base ($\mathcal{IGB}$) is proposed[15] as a set of valid and non redundant association rules. The $\mathcal{IGB}$ is the basis of Generic Association Rules, and is based on the Galois connection semantics. This generic basis is informative and compact[15]. The $\mathcal{IGB}$'s generic rules represent implications between minimal premises and maximal conclusions (in terms of the number of items). In fact, It was proven in the literature that this type of rules is the most general (i.e., conveying the maximum of information).

Most existing biclustering algorithms focus only on extracting positively-correlated genes. However, biological studies evince that a group of biologically significant genes may exhibit negative correlations. In the case of a positive correlation, genes present similar patterns while in a negative correlation, genes present opposite patterns (Figure 1). The problem of negatively-correlated patterns was studied in[16]. For example, the genes YLR367W and YKR057W of the Yeast dataset have a negative correlation with YML009C under 8 conditions. Worth mentioning, these genes are part of protein translation and translocation process, consequently, they should be grouped into the same cluster.

Several approaches are interested in extracting negative correlation biclusters[17,18,19,20,21]. This work deals with the problem of extracting biclusters with negative correlations from gene expression data using Generic Association Rules. The key of our NBic-ARM concerns the use of the $\mathcal{IGB}$ representation of the set of valid ARs defined by[15]. Our choice of this base is justified by the theoretical framework presented in[15].

To the best of our knowledge this is the first biclustering approach that employs the $\mathcal{IGB}$ representation to extract biclusters of negative correlations from gene expression data.

The remainder of the paper is organized as follows: Section 2 recalls the main definitions and notations that will be used throughout the remainder. Section 3 is dedicated to the description of our *NBic-ARM* algorithm. The encouraging results of the application of our algorithm on real-life microarray datasets are shown in Section 4. Conclusion and perspectives are sketched in Section 5.

2. Key notions

We give, in the following, the basic notions and definitions needed in this work;

Definition 1. ***(Biclustering)*** *A bicluster is a subset of objects (genes) associated with a subset of attributes (conditions) in which these rows are co-expressed. The bicluster associated with the matrix M =(I,J) is a couple (A,B), such that A ⊆ I and B ⊆ J, and (A,B) is maximal, i.e, if it does not exist a bicluster (C,D) with A ⊆ C or B ⊆ D.*
The biclustering problem focuses on the identification of the best biclusters of a given dataset. The best bicluster must fulfill a number of specific homogeneity and significance criteria (guaranteed through the use of a function to guide the search)[22].

Definition 2. ***(Formal context)*** *A formal context is a triple $\mathcal{K} = (G, M, \mathcal{R})$ where G is a set of objects, M is a set of attributes and the binary relation $\mathcal{R} \subseteq G \times M$ shows which objects have which attributes.*
A formal context can be represented by a cross- table. For $\mathcal{A} \subseteq G$, we define : $A^{'} = \{m \in M \mid \forall\ g\in \mathcal{A},\ (g, m) \in \mathcal{R}\}$ and dually for $\mathcal{B} \subseteq M$: $B^{'} = \{g \in G \mid \forall\ m\in \mathcal{B},\ (g, m) \in \mathcal{R}\}$. Roughly speaking, $A^{'}$ is the set of all attributes common to the objects of $\mathcal{A}$, while $B^{'}$ is the set of all objects that have all attributes in $\mathcal{B}$.

Definition 3. ***(Itemsets)*** *A non-empty finite set of $\mathcal{I} \subseteq M$ in $\mathcal{K}$ is called an itemset. An itemset containing k items is kalled $k-itemset$.*

Download English Version:

https://daneshyari.com/en/article/4960605

Download Persian Version:

https://daneshyari.com/article/4960605

Daneshyari.com