Information Sciences 415–416 (2017) 283–297

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Parallel biclustering detection using strength Pareto front evolutionary algorithm

Maryam Golchin, Alan Wee Chung Liew*

School of Information and Communication Technology, Gold Coast Campus, Griffith University, QLD 4222, Australia

ARTICLE INFO

Article history:
Received 19 December 2016
Revised 18 June 2017
Accepted 23 June 2017
Available online 26 June 2017

Keywords:
Biclustering
Gene expression dataset
Multi-objective evolutionary algorithm
Local search

ABSTRACT

Biclustering has become a popular technique to analyse gene expression datasets and extract valuable information by clustering rows and columns of a dataset simultaneously. Using a good merit function together with a suitable local search can lead to the detection of interesting biclusters. In this paper, a multi-objective evolutionary algorithm with local search is proposed to search for multiple biclusters concurrently in a single run of the evolutionary algorithm. We call our method PBD-SPEA2 (Parallel Biclustering Detection using Strength Pareto front Evolutionary Algorithm 2). In our algorithm, a new dynamic encoding scheme is used to encode multiple biclusters in each individual. Our multi-objective function consists of three objectives that simultaneously optimizes the homogeneity of the elements in the bicluster, the size of the bicluster, and the variance of the column in the bicluster with respect to the entire dataset. Crossover is done by selecting and combining the best biclusters among the encoded biclusters from both parents through a strategy of exploration and exploitation. Finally, a sequential selection procedure is used to select the final set of biclusters from individuals that constitute the Pareto front. Experimental results are presented to compare the performance and biological enrichment of detected biclusters with several existing algorithms.

1. Introduction

Microarrays measure the expression level of thousands of genes under various biological conditions. In a gene expression matrix, the rows typically describe gene annotations, and the columns represent conditions or sample annotations, as shown in Fig. 1. An $M \times N$ matrix represents gene expression dataset where M is the number of genes denoted by G and N is the number of conditions denoted by C. The matrix element e_{ij} is the gene expression value that shows the expression level of each gene under a particular condition. The genes behave similarly if they are co-regulated and share a similar function to a known pathway or structure. Sophisticated clustering algorithms that group genes into biologically meaningful groups based on their expression level have been widely used in genomic research, biomedicine, gene expression profiling, and discovery of diseases [12,36,37]. Clustering algorithms can be performed either row-wise or column-wise, and they compute similarity across the entire row or column profile. However, in many situations, a set of genes exhibit similar patterns only under a subset of conditions and would exhibit a different pattern under other conditions, and traditional clustering algorithms based on partitioned or hierarchical grouping do not have satisfactory performance on gene expression analysis [39]. In

* Corresponding author.
E-mail addresses: maryam.golchin@griffithuniversity.edu.au (M. Golchin), a.liew@griffith.edu.au (A.W.C. Liew).

http://dx.doi.org/10.1016/j.ins.2017.06.031
0020-0255/

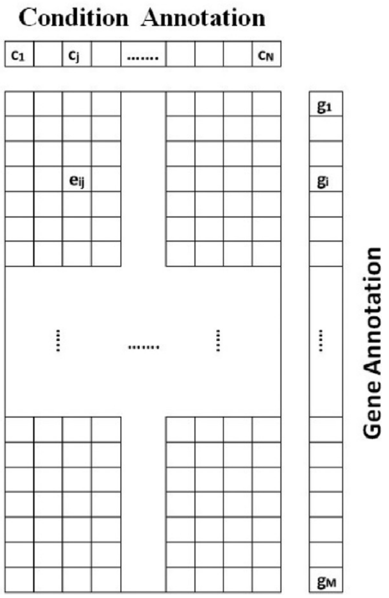

Fig. 1. An example of gene expression data.

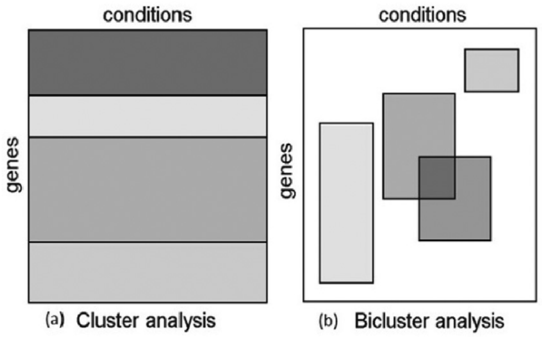

Fig. 2. Difference between (a) cluster analysis; and (b) bicluster analysis [39].

addition, most clustering techniques assign a given gene to only one cluster, whereas in reality a gene can participate in several different biochemical processes.

In order to discover the relationship between different genes and to avoid the drawbacks above, biclustering technique was proposed [3,8,9,11,13,15,17,20,22-28,32,34,39]. The goal of biclustering is to find subgroups of gene and conditions such that subset of conditions shows considerable homogeneity within a subset of genes to provide a better reflection of the biological reality. Biclustering refers to the clustering of both genes and conditions simultaneously to discover local patterns and similar transcriptional characteristics from microarray data. Here, a gene can take part in several biclusters under a different subset of conditions. Thus, biclusters can overlap. The difference between clustering and biclustering is visualized in Fig. 2.

Biclustering can be used to infer the biological role of an unknown gene by association with annotated genes in a bicluster. A recent survey of biclustering techniques can be found in [20]. The challenge of finding biclusters in datasets is an NP-hard problem [9] because the volume of the search space for finding biclusters increases exponentially when the number of genes and conditions increase.

The goal of this paper is to present a multi-objective evolutionary algorithm for bicluster analysis. Although many biclustering algorithms have been proposed, the problem remains largely unsolved. Here, we propose a biclustering approach based on the SPEA2 multi-objective evolutionary algorithm [40]. Unlike most existing multi-objective biclustering algorithms, we use a new dynamic encoding scheme to encode multiple biclusters in an individual. The new encoding scheme allows our algorithm to search for multiple biclusters in a dataset concurrently during each iteration of the evolutionary algorithm. We also introduced novel crossover and mutation operations to ensure individuals in the next generation are fitter than their parents. The rest of the paper is organized as follows. Section 2 provides an overview of evolutionary and