Efficient Mining Frequent Closed Discriminative Biclusters by Sample-Growth: The FDCluster Approach

Miao Wang, Northwestern Polytechnical University, China
Xuequn Shang, Northwestern Polytechnical University, China
Shaohua Zhang, Northwestern Polytechnical University, China
Zhanhuai Li, Northwestern Polytechnical University, China

ABSTRACT

DNA microarray technology has generated a large number of gene expression data. Biclustering is a methodology allowing for condition set and gene set points clustering simultaneously. It finds clusters of genes possessing similar characteristics together with biological conditions creating these similarities. Almost all the current biclustering algorithms find bicluster in one microarray dataset. In order to reduce the noise influence and find more biological biclusters, the authors propose the FDCluster algorithm in order to mine frequent closed discriminative bicluster in multiple microarray datasets. FDCluster uses Apriori property and several novel techniques for pruning to mine biclusters efficiently. To increase the space usage, FDCluster also utilizes several techniques to generate frequent closed bicluster without candidate maintenance in memory. The experimental results show that FDCluster is more effective than traditional methods in either single microarray dataset or multiple microarray datasets. This paper tests the biological significance using GO to show the proposed method is able to produce biologically relevant biclusters.

Keywords: Biclustering, DNA, Frequent Closed Discriminative Bicluster, Microarray, Weighted Undirected Sample Relational Graph

INTRODUCTION

Nowadays, in the post-genomic era, there have many bioinformatics data sets available. Due to the lack of accurate machine learning or intelligent tools in the bioinformatics community, the information embedded in most of these data has not yet completely exploited. Recently, DNA microarray technology has generated a large number of gene expression data, which is typically represented by a matrix where each cell represents the gene expression level of a gene under an experimental condition. How to use these data to reveal the function
and biological process of genes poses a great challenge of analysis algorithms. Various data mining techniques have been employed to infer useful biological information from the huge and rapid growing microarray data set.

One widely used method to infer relationship among genes in microarray data set is frequent pattern mining. Based on the characteristic of microarray data, (Pan et al., 2004; Cong et al., 2004) proposed to use condition enumeration method to exploit the gene patterns. However, both of above algorithms need to maintain the candidate patterns in memory, which limits the scalability. Association rules mining method is another way to analyze the gene expression data (Becquet et al., 2003; Creighton & Hanash, 2003; McIntosh & Chawla, 2007; Cong et al., 2004), which can discover the relationship among genes. However, it only can identify genes whose expression levels correlated across some conditions, it can not reveal the regulatory relations among genes. Using association rule to exploit regulatory modules has its limitations (Yeung et al., 2004).

How to identify genes with similar behavior with respect to different samples? Biclustering (Cheng & Church, 2000) is a methodology allowing for condition set and gene set points clustering simultaneously. It finds clusters of genes possessing similar characteristics together with biological conditions creating these similarities. The main advantage of biclustering is the simultaneous mining module on genes and experimental condition, another advantage is its applicability on original data instead of discretized data (Zhao & Zaki, 2005). However, mining microarray data for biclusters presents the following four challenges. First, the computing of biclustering method is NP-hard (Cheng & Church, 2000). Second, biclustering method deals with original data, it should adapt to the noise-sensitive character of microarray dataset. Third, the biclustering method should allow overlapping biclusters which share some genes or conditions, which would increase the complex of biclustering algorithm. Finally, the biclustering method should be flexible enough to handle different types of biclusters. (Madeira & Oliverira, 2004) classified biclusters into four categories: (i) constant value biclusters, (ii) constant row or column biclusters, (iii) biclusters with coherent values, where each row and column is obtained by addition or multiplication of the previous row and column by a constant value and (iv) biclusters with coherent evolutions, where the direction of change of values is important rather than the coherence of the values (Pandey et al., 2009).

Facing with the former three challenges above of biclustering method, some algorithms proposed to use greedy or heuristics approach for mining biclusters. In (Cheng & Church, 2000), Cheng and Church employed a greedy node deletion algorithm in their search based on using a low mean squared residue. Once a bicluster is created, its entries are replaced by random numbers and the procedure is repeated iteratively. Since then, there have been many greedy algorithms (Ben et al., 2003; Yang et al., 2003; Liu & Wang, 2007; Cheng et al., 2008; Teng & Chan, 2008; Dharan & Nair, 2009). A recent review of biclustering algorithms for biological data analysis can be found in (Madeira & Oliveira, 2004). Although these algorithms may improve the result, yet the efficiency is not very well. Another biclustering algorithm, MicroCluster (Zhao & Zaki, 2005), used weighted directed range multigraph to generate deterministic bicluster. The experimental results show this algorithm is very efficient. It also used some deletion or merging method to reduce the influence of noise. However, the main drawback of MicroCluster is that all the candidate biclusters need to be maintained in memory, which would increase computing complex and reduce space usage.

How to handle the last drawback of biclustering method is very intractable. Since the clustering formulation cannot be suitable to exploit some or all types of biclusters. For instance, SAMBA (Subramanian et al., 2005) is designed to find constant value biclusters. And Cheng and Church’s method (Cheng & Church, 2000) can find both constant value and constant row or column biclusters. Biclusters with coherent trends of up or down regulation
Combinatorial Fusion Analysis: Methods and Practices of Combining Multiple Scoring Systems
D. Frank Hsu, Yun-Sheng Chung and Bruce S. Kristal (2006). Advanced Data Mining Technologies in Bioinformatics (pp. 32-62).
www.igi-global.com/chapter/combinatorial-fusion-analysis/4245?camid=4v1a

Integrated Features Based on Gray-Level and Hu Moment-Invariants with Ant Colony System for Retinal Blood Vessels Segmentation
www.igi-global.com/article/integrated-features-based-gray-level/75154?camid=4v1a