Abstract

In a gene expression data matrix a bicluster is a sub-matrix of genes and conditions that exhibits a high correlation of expression activity across both rows and columns. The premise behind biclustering is that even related genes may only be expressed in a synchronized fashion over certain conditions. This approach overcomes some problems associated with traditional clustering methods, by allowing automatic discovery of similarity based on a subset of attributes, simultaneous clustering of genes and conditions, and overlapped grouping that provides a better representation for genes with multiple functions or regulated by many factors. Many biclustering algorithms rely on optimizing mean squared residue to discover biclusters from a gene expression dataset. In this thesis, a new biclustering method using a two-phase method has been proposed which can efficiently and accurately approximate k biclusters with low mean squared residue. The proposed approach is a two phase method of finding a bicluster. In the first phase, a modified version of *k-means* algorithm is applied to the gene expression matrix to generate clusters and check that the mean squared residue score of the clusters that are generated are within a threshold value. If the clusters that are formed have high residue score then the second phase of the algorithm is required where the columns are removed if the residue score of the columns are higher than the assigned parameter. Experimental study on yeast dataset shows that our algorithm finds better biclusters. Moreover, biological significance tests have been conducted to show that the biclusters identified using the proposed algorithm is composed of functionally enriched sets of genes.

Keywords: *biclustering, mean squared residue score, gene expression data, traditional clustering method*