ABSTRACT

It has been well established that gene expression data contain large amounts of random variation that affects both the analysis and the results of microarray experiments. Typically, microarray data are either tested for differential expression between conditions or grouped on the basis of profiles that are assessed temporally or across genetic or environmental conditions. Microarray technologies allow researchers to simultaneously monitor cellular activity of many gene transcripts. Clustering is often one of the first steps in Gene Expression Analysis. The goal of clustering is to subdivide the data (genes in our case) in such a way that similar items fall in the same cluster and dissimilar items fall in different clusters. This report presents a detailed and comparative view of the well-known clustering approaches with applications to gene expression data. A few of the clustering methods were experimented in light of real-life datasets and the methods have been established to perform satisfactorily. In this work, we propose an improved clustering algorithm that uses a modified Pearson's correlation measure to identify the clusters in gene expression data. Experimental results show the efficiency of the proposed method over several real-life datasets. The proposed method has been found to be better than other comparable algorithms in terms of z-score and p-value measures of cluster quality.

Keywords: Gene expression data; Microarray; Clustering; Pearson's correlation coefficient; z-score; p-value.