A COMPARISON OF CLUSTERING METHODS FOR MICROARRAY DATA BASED ON PREPROCESSING

S.Y. Kim, T. Hamasaki, T. Sugimoto

Osaka University Graduate School of Medicine, Osaka, Japan

Email: *gong@medstat.med.osaka-u.ac.jp*

Microarrays have become the effective, broadly used tools in biological and medical research to address a wide range of problems, including classification of disease subtypes or tumors. One of the major goals in analyzing microarray data is the detection of samples or genes with similar expression patterns. Due to the vast number of genes and the complexity of biological process, an effective clustering technique for grouping samples or genes is crucial to such studies. In medical cancer diagnoses based on microarray data, the definition of tumor classes would be based on clustering results. The results of clustering depend on various characteristics of the data, including the microarray experimental conditions, the variations between data points, and the degree of data noise. Data preprocessing is therefore an essential procedure for handling microarray data. In this study, we compare the performances of several clustering methods such as fuzzy c-means, k-means, partitioning around medoids and hierarchical clustering methods whilst considering data preprocessing by data normalization, effective gene selection among thousands of genes, and data noise. The performances of the clustering methods are compared using several simulated and real microarray datasets, with the results evaluated based on the validation measure. Consequently, we show that clustering methods which are common used in microarray data analysis are affected by data preprocessing.