Abstract

In data mining, preprocessing is one of important factor to yield efficient and good quality information. In unsupervised learning or clustering, the process of high dimension data will need expense and computing time that are big. Clustering process also can work better with data which have a little dimension.

The technique preprocessing, which is studied in this final duty, is Principal Component Analysis (PCA) where data set, which its dimension is big, summarized become data set with new dimension that its amount is slimmer. The new dimension is principal component (PC). PC formed by linear combination from original dimension so that data will not loss its genuiness characteristic.

Result of system examination of the colon data set tumor owning 2000 dimension can be summarized become 46 PC. PC and DLBCL data set owning 4026 dimension can be summarized become 46 PC. At data set the colon of tumor and DLBCL, data 1, 2, or 3 PC can give the performance result of K-Means Clustering which is better than the original data. For the method of Two Step Clustering of data set the colon tumor obtained by PCA performance which less be effective while the DLBCL data set obtained good performance of PCA at data 1 or 3 PC.

Keywords: data mining, preprocessing, PCA, clustering, multidimensi