Abstract

Gene expression analysis is getting better over years and so the amount of data produced increases. The range of different methods for processing data is also getting bigger each year. Clustering is an analysis method which deals with large datasets and tries to group the data with different distance measure methods. This approach leads to a better overview of the dataset and assists in creating results and drawing conclusions from the data.

Clustering is a good attempt to deal with big datasets and therefore it is often used in gene expression analysis. To get the best clustering results it is necessary to have a wide range of clustering possibilities like Hierarchical Clustering (HC) or k-means (KM) clustering.

In this thesis the clustering application Genesis was expanded by implementing the Mahalanobis Distance method, the Ward's linkage and two new clustering algorithms, Partitioning Around Medoids (PAM) and Transitive Clustering (TC).

To ensure correct implementation six data sets were used to check the new functions and features. Three datasets were taken from the already existing Genesis test files and were used to compare the clustering results from Hierarchical Clustering with different linkage measures and Mahalanobis distance measure with R and PAST. The remaining three datasets were used to validate the new clustering functions PAM and TC. These methods are using a precalculated similarity matrix to accelerate the clustering.

Comparable Silhouette value and F_1-score was yielded with different datasets. *Dataset 1* has a Silhouette value of 0.45 for TC and 0.4760 for PAM. The classification of the *Dataset 2* by TC achieves a F_1-score of 0,88 compared to 0.93 by ClustEval which is used as reference. PAM achieves a F_1-score of 0.91 by Genesis and 0.92 by ClustEval. The synthetic *Dataset 3* with 16 clusters has a F_1-score of 1.00 for TC and 0.76 for PAM. The *MCC* for *Dataset 3* is 1 for TC and 0.75 for PAM.

Key Words:

clustering, Genesis, Medoids, Transitive, Mahalanobis, Ward's