Abstract

Cluster analysis has been widely used over various domains to identify similar groups or patterns inherent in the data. The aim of this thesis is to study different clustering techniques applicable in spatial and gene expression data. For spatial data domain, this thesis presents three clustering techniques. The first technique (GDCT) is a grid-density based technique for identifying clusters of arbitrary shapes from 2D spatial data even in presence of noise. We succeeded in detecting the various types of outliers by using an outlier detection technique in-built with GDCT. The next two techniques (SATCLUS and GDSDC) are used to detect clusters in satellite data in a two phase process. Both SATCLUS and GDSDC use a grid-density based technique to identify the coarse clusters in the first phase. In the second phase, SATCLUS uses a partitioning strategy and GDSDC uses a fuzzy approach to obtain the final clusters from the coarse clusters. Also, to handle massive datasets, we propose two distributed techniques for 2D spatial data and satellite imagery. A study of clustering techniques for the analysis of gene expression data has also been discussed in this thesis. This thesis also includes two gene expression data clustering techniques. The first one is a density based clustering technique (GenClus) which clusters genes without taking the number of clusters as an input parameter. This thesis also incorporates an incremental version of the GenClus (InGenClus) to handle incremental gene expression data. The second technique GeneClusTree uses a hierarchical and density based approach to cluster the genes and forms a tree structure which helps in the visualization of the results. The proposed algorithm has been validated on several real-life datasets and found to perform well in comparison to similar algorithms. All clustering algorithms have been validated using various statistical measures.

Keywords -- Clustering, proximity measure, density based clustering, grid based clustering, embedded clusters, satellite data, coherent pattern, co-expressed gene, hierarchical clustering, distributed clustering