Two Phase Integrated Rule based Model (TPC-IRBM) for Clustering of Gene Expression Data of CA1 Region of Rat Hippocampus

Abstract

This paper propose a semi supervised clustering model TPC-IRBM (Two phase clustering-Integrated rule based model) for clustering large data set such as gene expression data. TPC-IRBM works in two phases to cluster the gene expression data set. The proposed model is based on rule based models CRT, C5, CHAID and QUEST. In the first phase of the model 30 % data (which may vary) is extracted to prepare training, testing and validation data (TTV data) using suitable heuristic or neural network based clustering techniques. The output of first phase is used as build the models and generate the rule base fitting to TTV data using aforesaid models. The proposed model is then constructed by selecting and integrating the quality rules of various models using qualifying criteria corresponding to every cluster. The number of quality rules in proposed model is much more compared to that of CRT, C5, CHAID and QUEST. The performance in terms of accuracy is better compared to the models. Although in some cases Neural Network based models performance is slightly better but a very high cost of complexity for very large data set.

References
Two Phase Integrated Rule based Model (TPC-IRBM) for Clustering of Gene Expression Data of CA1 Region of Rat Hippocampus


Statistica


Two Phase Integrated Rule based Model (TPC-IRBM) for Clustering of Gene Expression Data of CA1 Region

Index Terms

Computer Science  
Applied Sciences

Keywords

Gene expression clustering  semi supervised clustering  integrated rule based model  
two phase clustering  
CA1 region gene expression clustering of rat hippocampus.