Abstract

In data mining, clustering is one of the major tasks and aims at grouping the data objects into meaningful classes (clusters) such that the similarity of objects within clusters is maximized, and the similarity of objects between clusters is minimized. The dataset sometimes may be in mixed nature that is it may consist of both numeric and categorical type of data. Naturally these two types of data may differ in their characteristics. Due to the differences in their characteristics, in order to group these types of mixed data, it is better to use the ensemble clustering method which uses split and merge approach to solve this problem. In this paper, the original mixed dataset is splitted into numeric dataset and categorical dataset and clustered using both traditional clustering algorithms (K-Means and K-Modes) and fuzzy clustering algorithms (Fuzzy C-Means and Fuzzy C-Modes). The resultant clusters are combined using ensemble clustering methods and evaluated by both f-measure and entropy measure. It is found that splitting is more beneficial and applying fuzzy clustering algorithms yields better results than traditional clustering algorithms.

References

Ensemble Fuzzy Clustering for Mixed Numeric and Categorical Data

- Zengyou He, Xiaofei Xu, and Shengchun Deng "A Link Clustering based approach for Clustering Categorical Data", Department of Computer Science and Engineering, Harbin Institute of Technology, China.
- Roberto Avogadri and Giorgio Valentini, "Ensemble clustering with a fuzzy approach", Department of Science and Information (DSI), University of Milan, Italy.

Index Terms

Computer Science Data Mining
Keywords
Clustering  Ensemble Clustering  Mixed Dataset  Numeric Dataset  Categorical Dataset