Abstract

Categorical data has always posed a challenge in data analysis through clustering. With the increasing awareness about Big data analysis, the need for better clustering methods for categorical data and mixed data has arisen. The prevailing clustering algorithms are not suitable for clustering categorical data majorly because the distance functions used for continuous data are not applicable for categorical data. Recent research focuses on several different approaches for clustering categorical data. However, the complexity of methods makes them unsuitable for use in big data. Emphasis should be on algorithms which are faster. Thus paper proposes a simple, fast method derived from statistics for clustering categorical data. Results on popular datasets are encouraging.

References

A Simple Yet Fast Clustering Approach for Categorical Data

- D. Ienco, Ruggero G. Pensa and R. Meo, "Context-Based Distance Learning for Categorical Data Clustering," Advances in Intelligent Data Analysis VIII Lecture Notes in Computer Science Volume 5772, 2009, pp 83-94.
- Y Cheung and H Jia, "Categorical-and-numerical-attribute data clustering based on
a unified similarity metric without knowing cluster number"; Pattern Recognition 46 (2013) 2228–2238. Available at http://dx.doi.org/10.1016/j.patcog.2013.01.027

Index Terms

Computer Science
Database Management

Systems

Keywords

Clustering categorical data big data k-means