Abstract

Clustering is a method of finding homogeneous classes of the known objects. Clustering plays a major role in various applications in data mining such as, computational biology, medical diagnosis, information recovery, CRM, scientific data investigation, selling, and web analysis. Most of the researchers have a major interest in designing clustering algorithms. "Big data" involves terabytes and petabytes of data. Big data is challenging because of its five important characteristics such as volume, velocity, variety, variability and complexity. Therefore big data is difficult to handle using conventional tools and techniques. There are so many issues in clustering techniques, so some of the issues is how to process the data and big data is clustered in more compact format. Clustering algorithm suffer from stability problem, ensemble of single and multi level clustering. An important issue in clustering is that we do not have earlier knowledge regarding data. Also selection of input parameters such as number of
nearest neighbours, number of clusters in these algorithms makes clustering a challenging task. The main objective is to study and analyze the existing clustering algorithms, impact of dimensionality reduction and dealing with outliers.

References

- R. T. Ng and J. Han, Efficient and effective clustering methods for spatial data mining; in Proc. Int. Conf. Very Large Data Bases (VLDB), 1994, pp. 144–155
- G. Sheikholeslami, S. Chatterjee, and A. Zhang, Wavecluster: A multi resolution clustering approach for very large spatial databases; in Proc. Int. Conf. Very Large Data Bases (VLDB), 1998, pp. 428–439.
- A. Hinneburg and D. A. Keim, Optimal grid-clustering: Towards breaking the curse of dimensionality in high-dimensional clustering; in Proc. 25th Int. Conf. Very Large Data Bases (VLDB), 1999, pp. 506–517.

Index Terms

Computer Science

Information Sciences

Keywords

Clustering Algorithms Big Data Nearest Neighbours Outliers