Clustering in Big Data: A Review

Volume 153
Number 3

Year of Publication: 2016

Authors:
Anju, Preeti Gulia

Abstract

BIG DATA[1] is a term for data sets that are so large or complex that traditional data processing[4] applications are inadequate. Accuracy in big data may lead to more confident decision making, and better decisions can result in greater operational efficiency, cost reduction and reduced risk. Various algorithms and techniques like Classification, Clustering, Regression, Artificial Intelligence, Neural Networks, Association Rules, Decision Trees, Genetic Algorithm, Nearest Neighbor method are used for knowledge discovery from databases. Cluster is a group of objects that belongs to the same class. In other words, similar objects are grouped in one cluster and dissimilar objects are group in another cluster. Clustering methods can be classified into Partitioning Method, Hierarchical Method, Density-based Method. Clustering analysis is used in several applications like market research, pattern recognition, data analysis. K-means clustering is well known partitioning method. But this method has problem of empty cluster. The problems with existing system[6] were analysis, capture, search, sharing, storage, transfer, visualization, querying-updating. These problems can be reduced by using proposed algorithm. In this paper clustering and proposed algorithm is discussed.
References

7. Pei, Jian; Han, Jiawei; and Lakshmanan, Lak S. V.; Mining frequent itemsets with convertible constraints, in Proceedings of the 17th International Conference on Data Engineering, April 2–6, 2001, Heidelberg, Germany, 2001, pages 433-442
8. Agrawal, Rakesh; and Srikant, Ramakrishnan; Fast algorithms for mining association rules in large databases, in Bocca, Jorge B.; Jarke, Matthias; and Zaniolo, Carlo; editors, Proceedings of the 20th International Conference on Very Large Data Bases (VLDB), Santiago, Chile, September 1994, pages 487-499
10. Hájek, Petr; Havel, Ivan; Chytil, Metoděj; The GUHA method of automatic hypotheses determination, Computing 1 (1966) 293-308
15. Piatetsky-Shapiro, Gregory; Discovery, analysis, and presentation of strong rules, Knowledge Discovery in Databases, 1991, pp. 229-248
Index Terms

Computer Science
Distributed Systems

Keywords

Clustering, K-Mean, Data mining, Big data