STATISTICAL ANALYSIS AND DATA MINING

Research Article

A general framework for efficient clustering of large datasets based on activity detection*

Xin Jin, Sangkyum Kim, Jiawei Han, Liangliang Cao, Zhijun Yin

First published: 09 November 2010
https://doi.org/10.1002/sam.10097
Cited by: 1

* This work is extended from our SDM'09 conference paper [1]. Supported in part by the U.S. National Science Foundation grants IIS - 08 - 42769 and BDI - 05 - 15813 and IIS - 05 - 13678, and Office of Naval Research (ONR) grant N00014 - 08 - 1 - 0565. Any opinions, findings, and conclusions expressed here are those of the authors and do not necessarily reflect the views of the funding agencies.

Abstract

Data clustering is one of the most popular data mining techniques with broad applications. K - Means is one of the most popular clustering algorithms, due to its high efficiency/effectiveness and wide implementation in many commercial/noncommercial softwares. Performing efficient clustering on large dataset is especially useful; however, conducting K - Means clustering on large data suffers heavy computation burden which originates from the numerous distance calculations between the patterns and the centers. This paper proposes framework General Activity Detection (GAD) for fast clustering on large - scale data based on center activity detection. Within this framework, we design a set of algorithms for different scenarios: (i) exact GAD algorithm, E - GAD, which is much faster than K - Means and gets the same clustering result; (ii) approximate GAD algorithms with different assumptions, which are faster than E - GAD, while achieving different degrees of approximation; and (iii) GAD based algorithms to handle the *large clusters* problem which appears in many large - scale clustering applications. The framework provides a general solution to exploit activity detection for fast clustering in both exact and approximate scenarios, and our proposed algorithms within the framework can achieve very high speed. We have conducted extensive experiments on several datasets from various real world applications, including data compression, image clustering, and bioinformatics. By measuring the clustering quality and CPU time, the experiment results show the effectiveness and high efficiency of our proposed algorithms. Copyright © 2010 Wiley Periodicals, Inc. Statistical Analysis and Data Mining 4: 11–29 2011