INTISARI

EVALUASI KINERJA METODE DEKOMPOSISI REGULER UNTUK KLASTERISASI

Oleh

LAURA HARYO

20/466420/PPA/05986

Klasterisasi merupakan teknik pembelajaran mesin tanpa supervisi yang populer dipakai untuk analisis data di berbagai perusahaan besar. Perusahaan besar tentu memiliki data berjumlah besar pula yang perlu diolah. Dengan demikian, penting untuk memilih metode klasterisasi yang dapat bekerja dengan baik pada data berukuran besar.

Metode dekomposisi reguler merupakan salah satu metode klasterisasi yang diperkenalkan memiliki kemampuan untuk mengklasterisasi graf yang berukuran sangat besar. Namun, belum pernah diteliti lebih lanjut metode ini untuk kasus dataset numerik yang tidak langsung berpola graf. Selain itu, kedudukan metode ini juga belum pernah dikaji secara komprehensif untuk dataset berukuran sangat besar. Oleh karena itu, diajukan penelitian untuk mengevaluasi kinerja metode dekomposisi reguler dan membandingkannya dengan beberapa metode *benchmark* yang umum digunakan untuk klasterisasi data besar, yakni mini batch k-means dan DBSCAN, dan yang juga menggunakan konsep graf yakni spectral clustering.

Hasil penelitian menunjukkan bahwa RD menunjukkan akurasi yang tidak lebih baik dari spectral clustering dan mini batch k-means pada dataset berbentuk random dan terdistribusi normal, dengan selisih perbedaan maksimal mencapai 28,0%. Rata-rata konsumsi waktu dan memori metode RD juga selalu lebih besar daripada mini batch k-means dan DBSCAN. Nilai maksimal ARI yang dicapai RD pada 2.000 sampel KDD Cup 1999 hanya sebesar 40,7%, sangat rendah dibandingkan dengan DBSCAN yang dapat mencapai 96,7% pada jumlah sampel 50 kali lebih besar dan mini batch k-means yang dapat mencapai 74,1% pada jumlah sampel 2.449 kali lebih besar.

Kata Kunci: RD, Klasterisasi, Kinerja, *Big Data*, Perbandingan

ABSTRACT

PERFORMANCE EVALUATION OF REGULAR DECOMPOSITION FOR CLUSTERING

By

LAURA HARYO

20/466420/PPA/05986

Clustering is an unsupervised machine learning technique that is popularly used for data analysis in large companies. Big companies certainly have big data that needs to be processed. Thus, it is important to choose a clustering method that can work well on large data.

Regular decomposition is one of the clustering methods that was introduced for clustering very large graphs. However, this method has not been studied further for cases of numerical datasets that are not directly graph-patterned. In addition, the position of this method has not been studied comprehensively for very large datasets. Therefore, this research is proposed to comprehensively evaluate the performance of regular decomposition method and compare it with several benchmark methods commonly used to cluster big data, namely mini batch k-means and DBSCAN, and other method which also uses the graph concept, namely spectral clustering.

RD show no better accuracy than spectral clustering and mini batch k-means on normal-distributed dataset with random cluster, with a maximum difference of 28.0%. The average time and memory consumption of the RD method is also always greater than mini batch k-means and DBSCAN. The maximum ARI value achieved by RD in 2,000 samples of KDD Cup 1999 is 40.7%, very low compared to DBSCAN which can reach 96.7% at a sample size 50 times larger and mini batch k-means which can reach 74.1% with a sample size of 2,449 times larger.

Keywords: RD, Clustering, Performance, Big Data, Comparison