

INTISARI

ANALISIS KLASTER MENGGUNAKAN ALGORITMA CURE (CLUSTERING USING REPRESENTATIVES) UNTUK DATASET BESAR DENGAN PENCILAN

Oleh

Prima Lestari Hanif
14/364187/PA/15939

Metode *clustering* hierarki *agglomerative* merupakan metode analisis kluster yang pada mulanya menganggap setiap titik objek merupakan satu kluster sendiri, kemudian dikelompokkan berdasarkan kemiripan atau kesamaan karakteristik yang diukur menggunakan ukuran ketidaksamaan. Metode *clustering* hierarki *agglomerative* terbagi menjadi beberapa algoritma, diantaranya *Single Linkage*, *Complete Linkage*, *Average Linkage*, *Ward's Method*, dan *Centroid Method*. Namun, metode-metode ini sangat sensitif terhadap pencilan sehingga tidak dapat bekerja dengan baik dalam menangani kluster yang berbentuk *non spherical*. Oleh karena itu, dalam penelitian ini akan dibahas mengenai algoritma *Clustering Using Representatives* (CURE) yang mengkombinasikan pendekatan metode berbasis *centroid* dengan metode berbasis semua titik yang representatif.

Dalam prosesnya, CURE tidak menggunakan semua titik yang ada pada data melainkan hanya beberapa titik perwakilan yang cukup mewakili bentuk kluster dan ketersebaran data, titik ini dinamakan titik representatif. Sehingga, CURE efektif digunakan untuk dataset berukuran besar. Selain itu, CURE juga menyusutkan titik-titik data yang tersebar menuju pusat kluster menggunakan sebuah faktor penyusutan untuk mengurangi efek buruk karena pencilan. Untuk mengetahui kualitas dan validitas dari kluster, digunakan validasi internal *silhouette width*.

Kata kunci : analisis kluster, *Clustering Using Representatives*, faktor penyusutan, pencilan, dataset besar, *silhouette width*

ABSTRACT

CLUSTER ANALYSIS USING CURE (CLUSTERING USING REPRESENTATIVES) ALGORITHM FOR LARGE DATASET WITH OUTLIERS

by

Prima Lestari Hanif

14/364187/PA/15939

Agglomerative hierarchical clustering method is a cluster analysis method that initially considers each point of the object is a cluster, then grouped based on the similarity of characteristics that measured using dissimilarity measures. Some popular methods of agglomerative hierarchical clustering are Single Linkage, Complete Linkage, Average Linkage, Wards Method, and Centroid Method. However, these methods are very sensitive to outliers, so they cannot work properly in non-spherical clusters. Therefore, in this work we will discuss the Clustering Using Representatives (CURE) algorithm which combines the centroid-based method approach with all representatif point-based methods.

In the process, CURE does not use all the points of data but only a few points that can represent the shape of clusters and spread of data, this point is called a representative point. So, CURE is effectively used for large datasets. In addition, CURE also shrinks the scattered data points towards the center of the cluster using a shrinking factor to reduce the effect of outliers. To determine the quality and validity of the cluster, internal validation silhouette width was used.

Keywords: cluster analysis, Clustering Using Representatives, shrinking factor, outliers, large dataset, silhouette width