INTISARI

Algoritma *Weighted Robust Sparse K-Means (WRSK)* untuk Alternatif Analisis Klaster *Robust* dan *Sparse* yang Efisien pada Data Berdimensi Tinggi

oleh

Dita Dwi Aprilliani Ayu Lestari
14/364245/PA/15965

Minat yang semakin tinggi pada analisis data multivariat didorong oleh kebutuhan penelitian saat ini, yang mengarah pada sekumpulan peristiwa kompleks atau data berdimensi tinggi. Salah satu metode analisis data multivariat yang cukup populer adalah *K-Means*. Namun ketika dihadapkan pada data berdimensi tinggi, metode *K-Means* menjadi kurang optimal, mengingat semakin banyak permasalahan data yang muncul, diantaranya keberadaan *outliers* dan variabel mubazir (*noise*). Metode *K-Means* standar bersifat tidak *robust* karena menggunakan *mean* sebagai ukuran pusat klaster. Di samping itu, metode *K-Means* standar melibatkan keseluruhan variabel dalam proses *clustering*, sehingga sangat rentan terhadap pengaruh variabel mubazir (*noise*) yang mampu menyamarkan struktur klaster sesungguhnya. Data *real world* juga sering kali tidak memberikan informasi awal (prior), sehingga sulit melakukan estimasi terhadap parameter yang dibutuhkan dalam analisis klaster.

Untuk itu, pada skripsi ini membahas algoritma *Weighted Robust Sparse K-Means (WRSK)* sebagai metode analisis klaster untuk data berdimensi tinggi yang mampu melakukan identifikasi klaster, mendeteksi *outliers* serta menemukan variabel informatif secara bersamaan. Algoritma *WRSK* memuat tiga tahapan secara iterasi, yakni inisialisasi pusat klaster melalui metode ROBIN (*Robust Initialization*), deteksi *outliers* dengan pendekatan bobot (*weighted*) berdasarkan LOF (*Local Outliers Factor*), serta seleksi variabel berbasis algoritma *Sparse K-Means* dengan kendala $\|w\|_1 \leq s$ dan $\|w\|^2 \leq 1$. Dalam hal ini, estimasi terhadap parameter pusat awal maupun *outliers* dilakukan secara otomatis, sehingga *WRSK* dapat menjadi alternatif yang efisien bagi *RSK (Robust Sparse K-Means*). Selain itu, berdasarkan studi kasus yang dilakukan, diperoleh bahwa algoritma *WRSK* lebih unggul dalam melakukan analisis klaster pada data berdimensi tinggi.

Kata kunci*: K-Means, Robust Initialization, Local Outliers Factor,* Seleksi Variabel, *Weighted Robust Sparse K-Means,* Data Berdimensi Tinggi

ABSTRACT

Weighted Robust Sparse K-Means (WRSK) Algorithm for Efficient Alternative Robust and Sparse Clustering in High Dimensional Data

by

Dita Dwi Aprilliani Ayu Lestari
14/364245/PA/15965

Increasing of interest in multivariate data analysis is driven by current research needs, leading to a set of complex events or high-dimensional data. One of the most popular methods of multivariate data analysis is K-Means. However, when high dimensional data, K-Means method becomes less optimal, facing the number of data problems, containing both outliers and variable redundant (noise). The standard K-Means method is not robust because it uses the mean as the seed of the cluster. In addition, the standard K-Means method involves the entire variables in the clustering process, so it is particularly susceptible to the effect of redundant variables that can easily mask a real cluster structures. Real world data also often does not provide preliminary information, which make it difficult to estimate the parameters required in cluster analysis.

Therefore, this thesis discusses the Weighted Robust Sparse K-Means (WRSK) algorithm as a cluster analysis method for high-dimensional data that is able to identify clusters, detect outliers and find informative variables simultaneously. The WRSK algorithm contains three stages in iteration, i.e. cluster seed initialization through ROBIN (Robust Initialization) method, outliers detection with weighted approach based on LOF (Local Outlier Factor), and variable selection based on Sparse K-Means with constraints $\|w\|_1 \leq s$ *and* $\|w\|^2 \leq 1$. *In this case, estimation of the initial seed parameters and outliers are done automatically, so WRSK can be an efficient alternative to RSK (Robust Sparse K-Means). In addition, based on case studies, it was found that WRSK algorithm is superior in conducting cluster analysis on high-dimensional data.*

***Keywords**: K-Means, Robust Initialization, Local Outliers Factor, Variable Selection, Weighted Robust Sparse K-Means, High Dimensional Data*