INTISARI

PENGEMBANGAN METODE *ROBUST K-MEANS CLUSTERING* DAN APLIKASINYA

Oleh

ULFASARI RAFFLESIA

21/489810/SPA/00829

Clustering adalah metode dasar untuk mengelompokkan data berdasarkan kemiripan, dengan *k-means clustering* menjadi salah satu algoritma yang paling banyak digunakan karena kesederhanaan dan efisiensinya. Namun, algoritma *k-means clustering* standar sangat rentan terhadap pengaruh *outlier* atau pencilan, karena bergantung pada metrik jarak *non-robust* seperti jarak Euclidean. Mengatasi keterbatasan ini, dalam penelitian ini dikembangkan metode *robust k-means clustering*. Kontribusi pertama pada penelitian ini adalah kajian pendeteksian *outlier* pada data multivariat menggunakan metode jarak Mahalanobis dan deteksi pencilan spasial. Metode yang diamati dipilih karena efektivitasnya dalam menangani sifat multivariat dan ketergantungan spasial dari data, bersama dengan pendekatan *trimmed* dan *robust sparse k-means*.

Kajian kedua pada disertasi ini menghasilkan suatu bentuk perluasan baru dengan memodifikasi algoritma *k-means clustering* menggunakan beberapa metrik jarak baru yang *robust*, yaitu jarak *robust* Euclid standar dan jarak *robust* Mahalanobis. Fokus dari penelitian ini adalah pada kemampuan metode untuk mengelola efek *outlier* dan meningkatkan kinerja pengelompokan. Metode ini cukup efisien yang ditunjukkan oleh nilai kompleksitas waktu asimtitotis adalah $O(n)$.

Kemudian pada kajian ketiga, untuk memastikan efektivitas metode *k-means clustering*, indeks validitas klaster internal baru yang *robust* juga dikembangkan, mengatasi kelemahan indeks internal klasik yang rentan terhadap keberadaan *outlier*. Penelitian ini mengusulkan indeks validitas klaster berbasis median, *trimmed means*, *winsorized means*, *Huber mean* dan *MCD mean* sebagai alternatif pengganti *mean* yang tidak *robust* terhadap keberadaan *outlier*. Indeks-indeks ini bertujuan untuk menentukan jumlah klaster yang optimal dan meningkatkan keandalan evaluasi kualitas pengelompokan pada data dengan pencilan dengan memanfaatkan sejumlah indeks internal dalam metode *k-means clustering* yaitu Fukuyama-Sugeno Indeks (FS) dan Xie-beni Indeks (XB).

Metode-metode yang diusulkan dalam penelitian ini telah diterapkan pada

beberapa jenis data sebagai studi kasus. Aplikasi metode deteksi *outlier* multivariat berhasil mengidentifikasi sejumlah data yang menyimpang secara signifikan pada data gempa Sumatra, menunjukkan efektivitas pendekatan ini dalam mengenali pencilan. Selanjutnya, metode *k-means clustering* dan variannya mampu mengelompokkan data gempa ke dalam klaster yang merepresentasikan karakteristik tertentu, seperti magnitudo dan kedalaman. Penerapan metode *k-means clustering* berbasis jarak *robust* juga menunjukkan hasil yang unggul, di mana metrik seperti *Euclidean standar robust* dan *Mahalanobis robust* menghasilkan kualitas klaster yang lebih baik dan lebih tahan terhadap pengaruh pencilan, sebagaimana ditunjukkan oleh peningkatan nilai pada indeks validitas seperti Davies-Bouldin, Xie-Beni, dan Dunn. Terakhir, penentuan jumlah klaster optimal menggunakan indeks validitas klaster yang *robust* telah dilakukan pada data *iris*, *wine*, *breast cancer*, *pima Indian diabetes*, *glass identification*, *ecoli*, dan COVID-19, dengan hasil menunjukkan bahwa pendekatan berbasis MCD Mean memberikan hasil paling konsisten dan akurat dalam menentukan jumlah klaster optimal, baik pada data benchmark maupun data riil.

Dengan pengembangan baru yang telah diusulkan dalam disertasi ini, telah diberikan kontribusi pada bidang metode *clustering* yang lebih luas, dengan aplikasi pada berbagai bidang, terutama pada kejadian data yang rentan terhadap pencilan dan adanya variabilitas yang tinggi pada data.

Kata-kata kunci: *clustering*, *k-means*, *robust*, pencilan, ukuran jarak *robust*, validitas *robust*

ABSTRACT

DEVELOPMENT OF A ROBUST K-MEANS CLUSTERING METHOD AND ITS APPLICATION

By

ULFASARI RAFFLESIA

21/489810/SPA/00829

Clustering is a basic method for grouping data based on similarity, with k-means clustering being one of the most widely used algorithms due to its simplicity and efficiency. However, the standard k-means clustering algorithm is very vulnerable to the influence of outliers, as it relies on non-robust distance metrics such as Euclidean distance. To address this limitation, this research developed the robust k-means clustering method. The first contribution of this research is a study on outlier detection in multivariate data using the Mahalanobis distance method and spatial outlier detection. The observed methods were chosen for their effectiveness in handling the data's multivariate nature and spatial dependence, along with the trimmed and robust sparse k-means approaches.

The second study in this dissertation results in a new form of extension by modifying the k-means clustering algorithm using several new robust distance metrics, namely the standard robust Euclidean distance and the robust Mahalanobis distance. This research focuses on the method's ability to manage the effects of outliers and improve clustering performance. This method is quite efficient, as indicated by its asymptotic time complexity of $O(n)$.

Then, in the third study, to ensure the effectiveness of the k-means clustering method, a new robust internal cluster validity index was also developed, addressing the weaknesses of classical internal indices vulnerable to outliers. This study proposes cluster validity indices based on the median, trimmed means, winsorized means, Huber mean, and MCD mean as alternatives to the mean, which is not robust against the presence of outliers. These indices aim to determine the optimal number of clusters and enhance the reliability of clustering quality evaluation on data with outliers by utilizing several internal indices in the k-means clustering method, namely the Fukuyama-Sugeno Index (FS) and the Xie-Beni Index (XB).

The methods proposed in this study have been applied to several types of data as case studies. The application of the multivariate outlier detection method successfully identified several significantly deviating data points in the Sumatra

earthquake data, demonstrating the effectiveness of this approach in recognising outliers. Furthermore, the k-means clustering method and its variants are capable of grouping earthquake data into clusters that represent specific characteristics, such as magnitude and depth. The application of the robust distance-based k-means clustering method also shows superior results, where metrics such as robust standard Euclidean and robust Mahalanobis produce better cluster quality and are more resistant to the influence of outliers, as evidenced by the increased values on validity indices such as Davies-Bouldin, Xie-Beni, and Dunn. Finally, the determination of the optimal number of clusters using robust cluster validity indices has been conducted on the Iris, Wine, breast cancer, Pima Indian diabetes, glass identification, ecoli, and COVID-19 datasets, with results showing that the MCD Mean-based approach provides the most consistent and accurate results in determining the optimal number of clusters, both on benchmark and real-world data.

With the new developments proposed in this dissertation, contributions have been made to the broader field of clustering methods, with applications in various areas, especially in data occurrences that are prone to outliers and exhibit high variability.

Keywords: *clustering*, *k-means*, *robust*, *outlier*, *robust distance measure*, *robust validity*.