ABSTRACT

APPLICATION OF SILLHOUETTE COEFFICIENT METHOD, ELBOW METHOD, AND STATISTIC GAP METHOD IN DETERMINING OPTIMAL K IN K-MEDOIDS ANALYSIS

By

HILDA LAILATUL RAMADHANIA

Cluster analysis is a technique used for grouping data. One method of grouping data often used in cluster analysis is the non-hierarchical method. However, this method is weak in determining the number of clusters before analysis. One of the non-hierarchical methods that are often used is the K-Means method. This method is distance-based which divides the data into some clusters with numeric attributes. For data containing outliers, using the K-Means method is not recommended. Therefore, this method is modified so that it becomes the K-Medoids method. The difference between the two methods lies in selecting the medoid or the median value as the cluster's center. The thesis describes the research results using three methods to determine the optimal number of clusters on some data. The methods are the Sillhouette coefficient, the Elbow, and the Gap Statistics. According to the average Dunn Index value, the Gap Statistics method gave the largest one. Thus, the Gap Statistics method is recommended for research involving outlier data.

Key Words: *Sillhouette coefficient*, *Elbow*, *Gap Statistic*, *K-Medoids*.

ABSTRAK

APLIKASI METODE *SILLHOUETTE COEFFICIENT*, METODE *ELBOW*, DAN METODE *GAP STATISTIC* DALAM MENENTUKAN *K* OPTIMAL PADA ANALISIS *K-MEDOIDS*

Oleh

HILDA LAILATUL RAMADHANIA

Analisis klaster adalah teknik yang digunakan untuk mengelompokkan data. Salah satu metode pengelompokan data yang sering digunakan dalam analisis klaster adalah metode non-hierarki. Namun metode ini lemah dalam menentukan jumlah cluster sebelum dilakukan analisis. Salah satu metode non-hierarki yang sering digunakan adalah metode *K-Means*. Metode ini berbasis jarak yang membagi data menjadi beberapa cluster dengan atribut numerik. Untuk data yang mengandung *outlier*, tidak disarankan menggunakan metode *K-Means*. Oleh karena itu, metode ini dimodifikasi sehingga menjadi metode *K-Medoids*. Perbedaan kedua metode tersebut terletak pada pemilihan nilai *medoid* atau median sebagai pusat cluster. Skripsi ini memaparkan hasil penelitian dengan menggunakan tiga metode untuk menentukan jumlah cluster yang optimal pada beberapa data. Metodenya adalah koefisien *Sillhouette*, *Elbow*, dan *Gap Statistics*. Menurut nilai rata-rata *Dunn Index*, metode *Gap Statistics* memberikan yang terbesar. Oleh karena itu, metode *Gap Statistics* direkomendasikan untuk penelitian yang melibatkan data *outlier*.

Kata kunci: *Sillhouette coefficient, Elbow, Gap Statistic, K-Medoids.*