

INTISARI

Penentuan Jumlah Klaster Optimal Menggunakan Metode Gap Statistik

Oleh

Muftiani Solikhati Barokah

19/442598/PA/19347

Salah satu masalah utama dalam metode pengklasteran non-hierarki seperti *k-means* yakni harus ditentukan terlebih dahulu jumlah klaster yang hendak dibentuk. Metode gap statistik merupakan teknik yang dapat digunakan untuk menangani masalah tersebut dengan menentukan jumlah klaster optimal dalam pengklasteran. Gap statistik dapat digunakan dalam berbagai metode pengklasteran, tetapi dalam penelitian ini hanya difokuskan pada metode *k-means*. Gap statistik memperkenalkan nilai $Gap(k)$ yang merupakan hasil pengurangan nilai rata-rata logaritma *within cluster sum of square* yang direplikasi pada dataset yang telah dibangkitkan dengan nilai logaritma *within cluster sum of square* dari data asli. Studi kasus dilakukan terhadap data banyaknya desa/kelurahan menurut keberadaan dan jenis industri kecil dan mikro di tiap kabupaten/kota Provinsi Jawa Timur pada tahun 2018. Berdasarkan analisis yang dilakukan dengan metode gap statistik, diperoleh jumlah klaster optimal adalah sebanyak 4 klaster.

Kata Kunci: analisis klaster, *k-means*, gap statistik

ABSTRACT

Determining The Optimal Number of Cluster Using Gap Statistic Method

By

Muftiani Solikhati Barokah

19/442598/PA/19347

One of the main problems in non-hierarchical clustering methods such as k-means is that the number of clusters to be formed must be determined beforehand. The gap statistic method is a technique that can be used to deal with this problem by determining the optimal number of clusters in clustering. Gap statistic can be used in any clustering methods, but in this study is focused on the k-means method. The Gap statistic introduces the $Gap(k)$ value which is the result of subtracting the logarithm mean within cluster sum of square which is replicated in the dataset that has been generated with the logarithm value of within cluster sum of square from the actual data. Case studies were conducted on data on the number of villages according to the existence and type of small and micro industries in each district/city of East Java Province in 2018. Based on the analysis carried out with the gap statistic method, the optimal number of clusters was obtained as many as 4 clusters.

Keywords: clustering, k-means, gap statistic