INTISARI

PENGELOMPOKAN HIMPUNAN DATA MENGGUNAKAN METODE PARTISI *K-MEDOIDS* BERBASIS BLOK OBJEK

Oleh:
KARIYAM
21/475976/SPA/00777

Pada penelitian ini telah disusun algoritma pengelompokan untuk mempartisi n objek berdimensi p variabel dalam k grup. Algoritma yang diusulkan adalah metode partisi *K-Medoids* (KM) berbasis blok objek (Blok-KM atau BKM). *Medoid* adalah objek perwakilan suatu kelompok. Metode partisi Blok-KM terdiri atas dua tahap yaitu inisialisasi dan partisi. Inisialisasi adalah suatu tahap yang ditujukan untuk mendapatkan *medoid* awal, sedangkan partisi adalah suatu tahap untuk mendapatkan *medoid* akhir sebagai dasar mempartisi n objek.

Tahap inisialisasi dilakukan dengan mengambil perwakilan dari k blok objek terkecil pertama, dimana sebelumnya blok-blok objek telah disusun secara menaik. Blok objek yang dimaksudkan adalah sekumpulan objek yang mempunyai indikator standar deviasi sama dan jumlah data sama pada p variabel dengan atau tanpa standardisasi (BKM-O), ataupun pada n jarak antar objek (BKM-D), sehingga dalam n objek akan ditemukan sebanyak b blok objek $(b \leq n)$. *Medoid* awal yang dipilih adalah objek yang menempati urutan pertama dari setiap k blok objek $(k \leq b)$. Cara ini dapat menjamin bahwa objek identik yang terpilih sebagai *medoid* awal akan menempati kelompok awal dan kelompok akhir yang sama.

Tahap kedua metode partisi *k-medoids* berbasis blok objek menggunakan sejumlah kecil, (t), iterasi untuk menghasilkan *medoid* akhir yang stabil atau jumlah jarak dalam kelompok yang konstan. *Medoid* akhir adalah objek perwakilan kelompok yang mempunyai jumlah (atau ekuivalen dengan rata-rata) jarak objek ke *medoid* grup yang memuat objek adalah terkecil. Untuk melihat performa metode partisi Blok-KM digunakan *Rand Index, Adjusted Rand Index, Hubert Similarity Statistics, Jaccard Coefficient,* dan *Fowlkes-Mallows Index.* Capaian lima indek validasi eksternal dari metode partisi Blok-KM sebanding dengan metode pengelompokan di kelas partisi yang lain, khususnya metode *k-means* dan *simple and fast k-medoids* (SFKM). Metode partisi Blok-KM cukup efisien yang ditunjukkan oleh nilai kompleksitas waktu asimtitotis adalah $O(n^2)$.

Metode partisi Blok-KM, *k-means*, SFKM dan metode hirarki dengan tautan rata-rata, terpusat, tunggal, lengkap, rata-rata terbobot, serta metode Ward, secara umum akan dapat mengelompokkan himpunan data dengan lebih baik ketika dikerjakan pada data yang sebelumnya telah distandarkan, khususnya jika himpunan data terdiri atas beberapa variabel dengan nilai rentang data yang berbeda-beda baik dengan atau tanpa memuat data pencilan. Pada penelitian ini juga diusulkan teknik transformasi untuk standardisasi data. Transformasi data dikerjakan berdasarkan perbandingan antara data terhadap rentang data, dengan faktor pengali transformasi c, dimana data sebelumnya telah dirangking secara fraksional.

Pada penelitian ini juga dihasilkan kriteria indek rasio diviasi berbasis *medoid* (DRIM: *Deviation Ratio Index based on Medoids*). DRIM adalah suatu teknik untuk menentukan optimal cacah kelompok. Kriteria DRIM dikonstruksikan berdasarkan perbandingan indikator homogenitas dalam kelompok, $(SDW(k))$, terhadap indikator heterogenitas antar kelompok $(SDB(k))$. Kedua indikator ini dihitung berdasarkan jumlah jarak objek terhadap *medoid-medoid* kelompok akhir. Indikator homogenitas adalah jumlah jarak objek terhadap *medoid* kelompok yang memuat objek, sedangkan indikator heterogenitas dihitung berdasarkan jumlah jarak objek ke semua *medoid* grup lain yang tidak memuat objek. Validasi kriteria DRIM dilakukan dengan menggunakan indeks pembanding yang diusulkan oleh Calinski-Harabasz Kaufman-Rousseuuw, Hartigan dan Krzanowski-Lai. Tahap inisialisasi, teknik transformasi, dan kriteria untuk estimasi cacah kelompok ini, merupakan kontribusi utama dalam penelitian ini.

Kata kunci: partisi, kelompok, blok objek, *medoid*, validasi, homogenitas, heterogenitas, DRIM

ABSTRACT

CLUSTERING OF DATA SETS USING THE OBJECT BLOCK-BASED K-MEDOIDS PARTITIONING METHOD

By:
KARIYAM
21/475976/SPA/00777

This research has been developed a clustering algorithm to partition n objects with p variable dimensions into k groups. The proposed algorithm is an object block-based K-Medoids (KM) partitioning method (Block-KM). A medoid is a representative object of a group. The Block-KM partition method consists of two stages, namely initialization and partitioning. Initialization is a stage aimed at getting the initial medoid, while partitioning is a stage to get the final medoid as a basis for partitioning n objects. The initialization stage is carried out by taking representatives of the first k smallest object blocks, where previously, the object blocks have been arranged in ascending order. The block of objects in question is a group of objects that have the same standard deviation indicator and the sum of data on p variables with or without standardization (BKM-O) or at n distances between objects (BKM-D), so that in n objects there will be as many as b object blocks ($b \leq n$). The initial medoid chosen is the object that ranks first from each k block of objects ($k \leq b$). This method can guarantee that identical objects selected as initial medoids will occupy the same initial group and final group.

The second stage of the object block-based k-medoids partitioning method uses a small number (t) of iterations to produce a stable final medoid or a constant sum of distances within the group. The final medoid is a group representative object with the smallest total (or equivalent to the average) distance from the object to the group medoid containing the object. To validate the Block-KM partition method, Rand Statistics or Rand Index (RI), Adjusted Rand Index (ARI), Hubert Similarity Statistics, Jaccard Coefficient and Fowlkes-Mallows index. The external validation index achievements of the Block-KM partition method are comparable or relatively better with other clustering methods in partition class, especially for k-means and simple and fast k-medoids (SFKM). The Block-KM partition method is quite efficient as indicated by the asymtotic time complexity value of $O(n^2)$.

The Block-KM partition method, k-means, and hierarchical methods with the linkage of average, centroid, single and complete, McQuitty and Ward's method, will generally able to group data sets better when working on data that has previously been standardized, especially if the data set consists of several variables with different data range values, either with or without containing outlier data. In this research, a transformation technique for data standardization is also proposed. Data transformation is carried out by dividing the data by the data range, and multiplying it by the transformation multiplier factor c, where the previous data has been ranked fractionally.

This research also proposed a medoid-based deviation ratio index (DRIM: Deviation Ratio Index based on Medoids). The DRIM criteria is a technique for determining the optimal number of clusters. This criteria is constructed based on the ratio of the within-group

homogeneity indicator ($SDW(k)$), to the inter-group heterogeneity indicator ($SDB(k)$). These two indicators are calculated based on the total distance of the object to the final group of medoids. The homogeneity indicator is the sum of the object's distances to the group medoids contain the object, while the heterogeneity indicator is calculated based on the sum of the object's distances to all other group medoids that do not contain the object. Validation of the DRIM criteria was carried out using the comparison index proposed by Calinski-Harabasz, Kaufman-Rousseuuw, Hartigan and Krzanowski-Lai. The main contributions in this research are the initialization stage, the transformation technique, and the criteria for estimating the number of groups.

Keywords: partition, group, object-block, medoid, validation, homogeneity, heterogeneity, DRIM