Eduvest – Journal of Universal Studies
Volume 2 Number 10, October, 2022
p- ISSN 2775-3735- e-ISSN 2775-3727

# THE USE OF DISTANCE BLOCKS REPRESENTATIVE TO AVOID EMPTY GROUPS DUE TO NON-UNIQUE MEDOIDS

**Kariyam[1,2], Abdurakhman[2], Adhitya Ronnie Effendie[2]**
[1] Jurusan Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Islam Indonesia, Indonesia
[2] Jurusan Matematika, Fakultas MIPA, Universitas Gadjah Mada, Indonesia
Email: kariyam@uii.ac.id, rachmanstat@ugm.ac.id, adhityaronnie@ugm.ac.id

**ARTICLE INFO**

**ABSTRACT**

*The existence of identical objects in a data set is a necessity. This paper proposes a new indicator and procedure to obtain the initial medoids. The new algorithm guarantees no empty groups and identical objects in the same group, either in the initial or final groups. We use six real data sets to evaluate the proposed method and compare the results of other methods in terms of adjusted Rand index and clustering accuracy. The experiment results show that the performance of the proposed method is comparable with other methods.*

**KEYWORDS**

*clustering, identical objects, accuracy*



## INTRODUCTION

Research related to object grouping techniques and procedures has been carried out with various approaches; even with the development of big data, the need for grouping techniques has become inevitable. Jinchao, et al. (2021) proposed a multi-view clustering algorithm for mixed numeric and categorical data. Jinchao, et al. (2020) also proposed a mixed data clustering method based on Cuckoo Search and K-Prototypes. Behzadi, et al. (2020) investigated the use of the ClicoT (CLustering mix-type data including Concept Trees) for grouping mixed-type objects, especially nominal and numeric, and the method is noise-resistant grouping results can be interpreted well. This method can also automatically detect group enumeration even though there is no prior knowledge about the number of clusters. Selosse et al. (2019) have proposed using the Multiple Latent Block Model (MLBM) to classify mixed-type data sets using a stochastic algorithm approach. Oesting and Schnurr, (2020), have examined the relative position of an object in a group using

**How to cite:** Kariyam, Abdurakhman, Adhitya Ronnie Effendie (2022). The Use of Distance Blocks Representative To Avoid Empty Groups Due To Non-Unique Medoids. Journal Eduvest. *Vol 2* (10): Page 2218-2228
**E-ISSN:** 2775-3727
**Published by:** https://greenpublisher.id/

a group-size probability distribution approach on ordinal-type data.

Cluster analysis is generally distinguished into hierarchical, non-hierarchical and combination techniques (Jonhson & Wichern, 2009). One of the popular non-hierarchical techniques is the K-Means method. As the name implies, the K-Means method begins with determining k means that are determined directly or by partitioning objects into k groups and then calculating the cluster average. The K-Means method is unsuitable if there are outliers in the data set. To deal with this, the Partitioning Around Medoid (PAM) method, often called the K-Medoids, was proposed by Kauffman and Rousseeuw in 1987 and underwent development or modification. One problem of k-medoids is a high run time cost, so some investigations have been developed. Erich and Rousseuw (2021) have modified the second phase of k-medoid by eagerly performing additional swaps; so achieve an $O(k)$-fold speedup. At the same time, the purity algorithm has been developed to reduce the number of iterations by using of the Davies-Bouldin Index (Dinata et al., 2021). Nitesh et al. (2022) also improved k-medoids clustering based on the crow search.

A simple and fast k-medoids (Fast-KM) algorithm is one of the partitioning methods of datasets that consists of three phases (Park & Jun, 2009). The Fast-KM use a specific formula to select the set of initial medoids in the first stage. To update the medoids, the second phase of the Fast-KM use objects that minimize the total distance within groups. The third phase of Fast-KM is to assign objects to medoids. The second and third phases repeat until the total distance is stable. Even though this algorithm is simple and fast, it has neglected local optima and empty groups that may arise. For this reason, a simple k-medoids (Simple-KM) algorithm was developed to improve the performance of Fast-KM (Weksi & Leisch, 2019). The Simple-KM always use the most centrally-located object as one member of the initial medoids. Meanwhile, another medoid was determined randomly. The initialization repeats $s$ times. The Simple-KM suggestion for $s$ is twenty times. In case the medoids set are non-unique objects, for example, two medoids have equal values in all their variables, the Simple-KM regulate that the closest objects to the non-unique medoids assign only one of the medoids. In the Simple-KM, to avoid an empty cluster, the non-unique medoids are restricted to preserve their label, meaning that identical objects may occur in different groups during the initialization process to obtain initial groups. The flexible k-medoids (Flexible-KM) have been developed to improve the performance of Fast-KM and Simple-KM (Kariyam et al., 2022). The Flexible-KM use an object representing a block of a combination of standard deviation and the sum of variable values. This algorithm guarantee that no empty groups and identical object as non-unique objects are in the same group in the initialization process.

This research aims to develop another way to improve the performance of the Fast-KM addressed to avoid the empty group due to non-unique medoids. We apply the proposed method to the six real datasets from the University of California, Irvine (UCI) repository to evaluate it.

## RESEARCH METHOD

### Proximity measure and transformation method

We use a simple matching coefficient to get the matrix distance for binary, categorical, and nominal variables (Everitt et al., 2011). For numerical data, we apply Euclidean or Manhattan distance. While for mixed data (categorical and numeric), we implement a Generalized Distance Function (GDF). The GDF measure between objects $i$ and $j$ for non-missing mixed data as follows (Kauffman & Rousseuw, 1990),

$$d_{ij} = \sum_{s=1}^{p_b} \delta_b(x_{is}, x_{js}) + \sum_{t=1}^{p_c} \delta_c(x_{it}, x_{jt}) + \sum_{r=1}^{p_n} \delta_n(x_{ir}, x_{jr}). \quad (1)$$

where $p_b$, $p_c$, and $p_n$ are the number of binary, categorical, and numeric variables, respectively. The simple matching distance between objects $i$ and $j$ for the variable $f$ are as follows (Xu & Wuncsh, 2009),

$$d_{ij}^f = \begin{cases} 1 & if \ x_{if} \neq x_{jf} \\ 0 & if \ x_{if} = x_{jf} \end{cases} \quad (2)$$

The Euclidean distance between objects $i$ and $j$ are defined as follows,

$$d_{ij} = \left[\sum_{l=1}^{p} (x_{il} - x_{jl})^2\right]^{1/2} \quad (3)$$

We can standardize numerical data if the datasets are mixed or contain outliers (Jajuga, 2000). The transformations on numerical data are linear transformations with a standardization formula as follows [11],

$$z_{ij} = bx_{ij} + a \quad (b > 0), \quad (4)$$

where $x_{ij}(z_{ij})$ denotes the value (standardized value) of the $j$th variable for $i$th object. The value is often used is $b = \frac{1}{\sigma}$ and $a = -\frac{\mu}{\sigma}$. The transformation for the ratio scale also uses the value of $b = \frac{1}{x_{0j}}$ and $a = 0$, where $x_{0j}$ denotes normalizing value, depending on cases that are met, for example, range, the maximum value of a variable, standard deviation, or mean. Meanwhile, the rank-based transformation is used for ordinal data (Zorn, 2003). This paper use standardized method to avoid an influence from the attribute values dimension using Eq. (5) below (Kariyam et al., 2022),

$$Z_{li} = f.\left(\frac{r_{li} - r_{l1}}{r_{lm} - r_{l1}}\right); \quad i = 1, 2, \ldots, n \quad (5)$$

where $r_{li}$ is the rank of object $ith$ variable $l$, and $f$ is the transformation multiplier for standardization.

**External validation measures**

We use Purity and F-measure to validation of our method. Table 1 shows the contingency matrix, namely, the crosstab between the members of the true partition in the data set (partition C) and separation based on a clustering method (partition D) (Wu, 2012).

The formula of Purity is as follows,

$$P = \sum_{i=1}^{k} p_i \left(max_j \frac{p_{ij}}{p_i}\right). \quad (6)$$

The formula of the F-measure is as follows,

$$F = \sum_{j=1}^{k'} p_j \, max_i \left( 2 \frac{p_{ij}}{p_i} \frac{p_{ij}}{p_j} / \left( \frac{p_{ij}}{p_i} + \frac{p_{ij}}{p_j} \right) \right). \quad (7)$$

**Table 1**
**The contingency matrix**

| | | Partition C | | | | |
|---|---|---|---|---|---|---|
| | | $C_1$ | $C_2$ | ⋯ | $C_{k'}$ | Σ |
| Partition P | $P_1$ | $n_{11}$ | $n_{12}$ | ⋯ | $n_{k'1}$ | $n_{1.}$ |
| | $P_2$ | $n_{21}$ | $n_{22}$ | ⋯ | $n_{k'2}$ | $n_{2.}$ |
| | . | . | . | ⋯ | . | . |
| | $P_k$ | $n_{k1}$ | $n_{k2}$ | ⋯ | $n_{kk'}$ | $n_{k.}$ |
| | $\Sigma$ | $n_{.1}$ | $n_{.2}$ | ⋯ | $n_{.k'}$ | $n$ |

Meanwhile, the formula for the accuracy level is as follows,

$$Acc = \frac{n_{11}+n_{22}+\cdots+n_{kk'}}{n} \; x \; 100\%. \quad (8)$$

We also use the adjusted Rand index (ARI) to evaluate the proposed method. Suppose that C is a clustering result under consideration and P is the actual partition, then the ARI is formulated as follows (Warrens & Hoef, 2022),

$$ARI = \frac{2(ad-bc)}{(a+b)(b+d)+(a+c)(c+d)}, \quad (9)$$

where $a$ is the number of pairs of the objects placed in the same cluster in P and the same group in C; $b$ is the number of pairs in the same class in P but not in the same cluster in C, $c$ is the number of pairs in the same group in C but not in the same cluster in P, and $d$ is the number of pairs in different groups in C and different classes in P.

## Proposed Method

A simple and fast k-medoids algorithm selects the initial medoids using a set of ordered $v_j$. The value of $v_j$ is calculated based on standardized row sum or standardized column sum of the distance matrix, such as follows (Park & Jun, 2009),

$$v_j = \sum_{i=1}^{n} \frac{d_{ji}}{\sum_{l=1}^{n} d_{il}}, \quad i = j = l = 1, 2, \ldots, n. \quad (6)$$

The initial medoids were chosen based on the first $k$ smallest set from the ordered $v_j$. On the other hand, the flexible k-medoids use a representative object of a block of combination standard deviation and sum of variable values as the initial medoids.

This paper considers an empty group that may occur in simple and fast k-medoids due to non-unique medoids. We adopt the idea of flexible k-medoids to obtain the initial groups, using the combination of standard deviation (variance) and a sum of p-variables values. We call this method a k-medoids algorithm based on distance variance blocks (Var-KM). The detail of our proposed algorithm is as bellows.

**Stage 1: (Determine the initial groups)**

(i) Calculate the distance matrix for all pairs of objects, $D_{ij}, (i = j =$

$1, 2, \cdots, n)$

(ii) For each object $i$, $(i = 1, 2, \cdots, n)$, calculated two parameters, namely the sum of distances $(w_i)$ and variance of distances $(u_i)$, such as follows:

$$w_i = \sum_{j=1}^{n} d_{ij}, \tag{7}$$

$$u_i = \frac{\sum_{j=1}^{n} (d_{ij} - \bar{d}_i)^2}{n-1}, \tag{8}$$

where $\bar{d}_i = \frac{w_i}{n}$.

(iii) Arrange all objects, first based on standard deviation such as Equation (8), $u_j$, in ascending order, then each block of the same value of $u_i$ (if any), objects are sorted based on the sum of distances, such as Equation (7), $w_i$, also in ascending order.

(iv) For the first $k$ blocks of the combination of $u_i$ and $w_i$ (or may only block of $u_i$); select the first object from each block as the initial medoid.

(v) Determine the members of $k$ initial groups based on the distance of an object to the nearest medoid.

Stage 2: (Determine the final groups)

(i) Update the current medoid in each cluster based on the object that minimizes the average distance to other things in its group. The average distance within cluster $g$th, which has $n_g$ members for object $i$th, $\overline{D}_i$, defined as follows,

$$\overline{D}_i = \frac{1}{n_g} \sum_{j=1}^{n_g} d_{ij}. \tag{11}$$

(ii) Obtain the cluster by assigning each object to the nearest medoid and calculate the total distance from all items to their medoids, $TD(k)$, such as follows,

$$TD(k) = \sum_{g=1}^{k} \sum_{i=1}^{n_g} \sum_{l=1}^{p} \left| x_{gil} - m_{gl} \right|, \tag{12}$$

where $x_{gil}$ is object $i$th for variable $l$th in the cluster $g$th; and $m_{gl}$ is medoid cluster $g$th for variable $l$th.

(iii) Repeat steps (i) and (ii) until the $TD(k)$ is equal to the previous one or a pre-determined number of iterations is attained, or the set of medoids does not change.

The novelty of Var-KM (proposed method) is the step for determining the initial groups. We use different indicators to search the initial medoid, such as steps (ii), (iii) and (iv) in the first stage. We combine the two and three phases of the Fast-KM into one scene. In addition, we add the regulation to obtain the final groups using a pre-determined number of iterations to achieve stability of total distance or set of medoids does not change. These processes ensure that there are no empty groups and that the identical objects or non-unique medoids are in the same group, either in the initial or final group. The source to calculate indicators in the Var-KM

differs from the flexible k-medoids algorithm. The flexible k-medoids directly calculated the standard deviation from the original or standardized data. The steps of the second phase of the Var-KM are also different from the flexible k-medoids. In the second phase of the Flexible-KM, use B times randomly to obtain the final medoids, meaning that different random sizes could produce no similar member groups.

**An illustrative example of k-medoids based on distance variance blocks**

Suppose we have fourteen objects observed on three numerical variables, such as in Table 2.

**Table 2**
**Artificial small data**

| Variables | Objects | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **A** | **B** | **C** | **D** | **E** | **F** | **G** | **H** | **I** | **J** | **K** | **L** | **M** | **N** |
| X | 35 | 65 | 70 | 45 | 45 | 20 | 25 | 10 | 30 | 15 | 100 | 90 | 85 | 96 |
| Y | 10 | 15 | 40 | 45 | 45 | 70 | 80 | 65 | 70 | 70 | 80 | 85 | 90 | 92 |
| Z | 15 | 20 | 15 | 20 | 20 | 50 | 45 | 50 | 55 | 55 | 20 | 25 | 35 | 30 |

From table 2., we get the Euclidean distance matrix, such as follows,

$$D_{ij} = \begin{bmatrix} 0 & 30.8 & 46.1 & 36.7 & 36.7 & 71.1 & 76.8 & 69.8 & 72.3 & 74.8 & 95.7 & 93.5 & 96.4 & 103.3 \\ 30.8 & 0 & 25.9 & 36,1 & 36.1 & 77.1 & 80.3 & 80.2 & 73.9 & 82.2 & 73.8 & 74.5 & 79.1 & 83.6 \\ 46.1 & 25.9 & 0 & 25.9 & 25.9 & 68.0 & 67.3 & 73.8 & 64.0 & 74.3 & 50.2 & 50.2 & 55.9 & 60.0 \\ 36.7 & 36.1 & 25.9 & 0 & 0.0 & 46.3 & 47.4 & 50.2 & 45.6 & 52.4 & 65.2 & 60.4 & 62.0 & 70.1 \\ 36.7 & 36.1 & 25.9 & 0.0 & 0 & 46.3 & 47.4 & 50.2 & 45.6 & 52.4 & 65.2 & 60.4 & 62.0 & 70.1 \\ 71.1 & 77.1 & 68.0 & 46.3 & 46.3 & 0 & 12.2 & 11.2 & 11.2 & 7.1 & 86.0 & 75.8 & 69.6 & 81.6 \\ 76.8 & 80.3 & 67.3 & 47.4 & 47.4 & 12.2 & 0 & 21.8 & 15.0 & 17.3 & 79.1 & 68.2 & 61.6 & 73.6 \\ 69.8 & 80.2 & 73.8 & 50.2 & 50.2 & 11.2 & 21.8 & 0 & 21.2 & 8.7 & 96.0 & 86.2 & 80.5 & 92.3 \\ 72.3 & 73.9 & 64.0 & 45.6 & 45.6 & 11.2 & 15.0 & 21.2 & 0 & 15.0 & 78.9 & 68.7 & 61.8 & 73.9 \\ 74.8 & 82.2 & 74.3 & 52.4 & 52.4 & 7.1 & 17.3 & 8.7 & 15.0 & 0 & 92.5 & 82.2 & 75.5 & 87.6 \\ 95.7 & 73.8 & 50.2 & 65.2 & 65.2 & 86.0 & 79.1 & 96.0 & 78.9 & 92.5 & 0 & 12.2 & 23.5 & 16.1 \\ 93.5 & 74.5 & 50.2 & 60.4 & 60.4 & 75.8 & 68.2 & 86.2 & 68.7 & 82.2 & 12.2 & 0 & 12.2 & 10.5 \\ 96.4 & 79.1 & 55.9 & 62.0 & 62.0 & 69.6 & 61.6 & 80.5 & 61.8 & 75.5 & 23.5 & 12.2 & 0 & 12.2 \\ 103.3 & 83.6 & 60.0 & 70.1 & 70.1 & 81.6 & 73.6 & 92.3 & 73.9 & 87.6 & 16.1 & 10.5 & 12.2 & 0 \end{bmatrix}$$

We clustered datasets into three groups. If we implement fast k-medoids algorithms based on indicator $v_j$ that calculated according to Equation (6), then we obtained objects D, E, and I as initial medoids such as columns (2) in Table 3. These initial medoids contain non-unique objects, namely objects D and E (both identical entities), which is caused by the fast k-medoids algorithm failing to group into three groups. A fast k-medoids algorithm grouped datasets into two groups.

Table 3. Artificial small data, for example of the proposed method

| **Fast-KM** | | **Proposed methods** | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | **First phase** | | | | **Second phase** | | |
| | | | | | | **Iteration (I)** | | |
| **Objects ordered** | $v_j$ | **Objects ordered** | **Std. dev. of distance ($u_i$)** | **Sum of distance ($w_i$)** | **Initial groups** | **I-1** | **I-2** | **I-3**** |

| (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
|---|---|---|---|---|---|---|---|---|
| D | 0,798* | D | 21,741 | 598,550* | G1* | G3 | G3* | G3* |
| E | 0,798* | E | 21,741 | 598,550 | G1 | G3 | G3 | G3 |
| I | 0,859* | C | 22,102 | 687,943* | G2* | G3 | G3 | G3 |
| F | 0,879 | B | 27,559 | 833,655* | G3* | G3 | G3 | G3 |
| G | 0,888 | I | 28,140 | 647,216 | G1 | G1 | G1 | G1 |
| C | 0,946 | G | 28,764 | 668,067 | G1 | G1 | G1 | G1 |
| J | 0,961 | M | 29,644 | 752,537 | G2 | G2 | G2 | G2 |
| H | 0,989 | A | 30,259 | 904,147 | G3 | G3* | G3 | G3 |
| M | 1,055 | L | 31,743 | 755,188 | G2 | G2* | G2* | G2* |
| L | 1,064 | F | 32,307 | 663,725 | G1 | G1* | G1* | G1* |
| B | 1,157 | K | 33,399 | 834,429 | G2 | G2 | G2 | G2 |
| N | 1.178 | H | 34,229 | 742,162 | G1 | G1 | G1 | G1 |
| K | 1,179 | J | 34,491 | 721,956 | G1 | G1 | G1 | G1 |
| A | 1,249 | N | 34,574 | 834,943 | G2 | G2 | G2 | G2 |
| | | Total deviation of groups, $TD(k)$ | | | 489,3 | 227,1 | 175,4 | 175,4 |

* Medoids ** Final groups

Meanwhile, in the first phase of Var-KM (proposed method), objects D, C and B are initial medoids such as columns (5) in Table 3. These medoids resulted from the initial groups, such as columns (6) in Table 3. We can see that the identical object, i.e. D and E, are in the same group. In order, the second phase of the proposed method concluded that the total deviation of the third iteration is equal to the second iteration, such as columns (8) and (9). Even though the group members of iteration one are similar to iterations two and three, the total deviation of iteration one is more excellent than iteration two, so we concluded that the final groups are iteration two (equal to iteration three). The profile of the initial groups and final groups of the proposed method are in Figure 1a and Figure 1b.

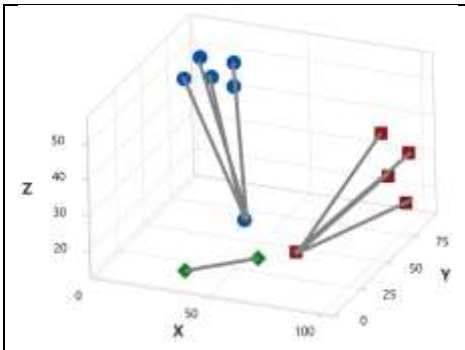


Figure 1a. The plot of the initial groups

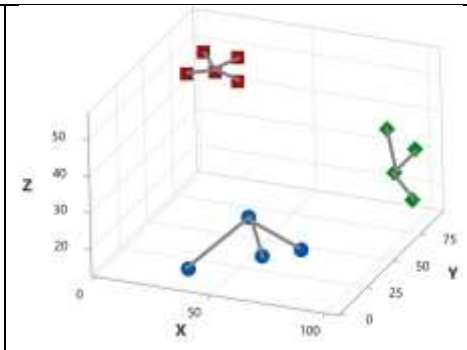


Figure 1a. The plot of the final groups

## Real datasets

Six datasets from the University of California, Irvine repository, namely ionosphere, breast cancer, primary tumour, vote, soybean small, zoo, and credit approval, are applied to check the effectiveness of our proposed methods (Lichman, 2022). The profile of six datasets, such as Table 4.

**Table 4**
**Profile of the real datasets**

| Data Set | $n$ | $p_n$ | $p_c$ | $k$ | Type |
|---|---|---|---|---|---|
| 1. Breast cancer | 683 | 9 | - | 2 | Numerical |
| 2. Ionosphere | 351 | 33 | - | 2 | Numerical |
| 3. Primary tumour | 65 | - | 17 | 4 | Categorical |
| 4. Soybean small | 47 | - | 35 | 4 | Categorical |
| 5. Zoo | 101 | 1 | 15 | 7 | Mixed |
| 6. Heart disease (HD) case 2 | 303 | 5 | 8 | 2 | Mixed |

$n$: number of objects; $p_n$: number of numerical variables; $p_c$: number of categorical variables; $k$: number of actual clusters

## RESULT AND DISCUSSION

From the Fast-KM, Var-KM (proposed), and other methods, we obtained the comparison of adjusted Rand index (ARI) and clustering accuracy, such as in Table 5.

The breast cancer dataset involves nine numerical variables computed from a digitized image describing the cell nuclei characteristics. These data are classified into two classes and consist of 683 non-missing cases. We apply Equation (5) for all features with f=5 and use Euclidean distance. The fast k-medoids algorithm fails to group this data into two classes because the medoids are non-unique objects. Breast cancer data has many identical objects, so if the initialization is random, it is possible to find empty groups. A k-medoids algorithm based on distance variance blocks (Var-KM) can handle it well. The adjusted Rand index and accuracy of Var-KM (proposed method) are comparable with methods offered by Ji et al., 2020 and Ji et al., 2021. The ionosphere data contains 351 objects with 34 numerical features assigned into two groups. The ARI and accuracy of the Fast k-medoids algorithm are similar to the proposed method.

**Table 5**
**The comparison of ARI and accuracy from the Var-KM, Fast-KM and others**

| Datasets | Adjusted Rand index | | | Clustering accuracy (in per cent) | | |
|---|---|---|---|---|---|---|
| | Var | Fast | Others | Var | Fast | Others |
| 1 Breast cancer | 0.869 | Fail | 0.891[a] | 96.6 | Fail | 97.2[a], 95.8[d], 95.6[e] |
| 2 Ionosphere | 0.173 | 0.146 | 0.292[a], 0.173[c] | 70.9 | 69.2 | 78.4[a] |
| 3 Primary tumour | 0.181 | Fail | 0.336[a] | 50.8 | Fail | 68.2[a] |
| 4 Soybean small | 1.000 | 0.498 | 1.000[a, b, c] | 100 | 80.9 | 100[a,b], 97.8[f], 98.5[d] |
| 5 Zoo | 0.963 | Fail | 0.967[a], 0.901[d], 0.939[e] | 92.1 | Fail | 96[a], 82.2[b], 88.8[d], 89.9[e] |
| 6 HD case 2 | 0.360 | 0.16 | 0.430[a] | 80.2 | 70.3 | 84.2[a], 81.2[d], 81.0[e] |

[a](Kariyam et al., 2022), [b](Weksi & Leisch, 2019), [c](Yu et al., 2018), [d](Ji, et al., 2021), [e](Ji, et al., 2020), [f](Yuan et al., 2020)

As the breast cancer dataset, the Fast-KM fails to cluster the primary tumour into four groups. The Fast-KM algorithm produces two empty initial groups caused by the second, third and fourth medoids being identical objects. Even though the ARI and accuracy of Var-KM are not very high, this method can avoid empty groups on Fast-KM. The soybean small dataset has 35 variables; three have an ordinal type tendency, namely precip, temperature and germination features. This data consists of 47 items assigned to four classes. We operate simple matching distance for categorical data. We utilize Equation (5) with f=1 for ordinal data before computing the Manhattan distance. The adjusted Rand index and accuracy of the proposed method are perfectly one or hundred per cent. This achievement is equal to flexible k-medoids (Kariyam et al., 2022) and simple k-medoids (Weksi & Leisch, 2019). In comparison, the Fast-KM produces an accuracy of 80.9% with an ARI of 0.498. Meanwhile, the clustering accuracy by Ji et al. (2021) and Yuan et al. (2020) are less than the proposed method.

The zoo dataset has 101 animals assigned to seven clusters, comprising 15 binary attributes and one numerical data. We use Equation (5) with value factor f=1 for numerical data before executing Manhattan distance. The zoo dataset has 25 blocks of objects with the same standard deviation and 30 object blocks with the same standard deviation and sums up all variables. If the Fast-KM method is used, two initial clusters will empty because the medoid groups are non-unique. The Var-KM produces the adjusted Rand index and accuracy with high values and is comparable to other methods. For the last real data, i.e. the heart disease case 2 data, the ARI and accuracy of the proposed method are higher than fast k-medoids. The heart disease case 2 data comprised 303 patients assigned into two clusters. These data are mixed variables with three binary, five categorical and five numerical data. Before implementing the Manhattan distance, we applied Equation (5) for numerical data, using the factor of f=5. Meanwhile, we take simple matching for others.

Based on Table 5, we concluded that the k-medoids algorithm based on distance variance blocks could avoid empty groups that may occur in the Fast-KM. The adjusted Rand index and clustering accuracy of Var-KM for six real datasets are generally relatively comparable to other methods.

To balance the evaluation of our proposed method, we calculate Purity and F-measure for six real datasets, such as in Table 6. According to Table 6, the k-medoids algorithm based on distance variance blocks produces Purity and F-measure values of more than 0.7, especially for breast cancer, ionosphere, soybean small and heart disease case 2 data datasets. Even though the Purity and F-measure of the primary tumour and zoo datasets are unsatisfaction, the new method can cluster into a certain number of groups and no empty classes.

**Table 6**
**External validation for the proposed method**

| Datasets | Purity | F-measure |
|---|---|---|

| | | | |
|---|---|---|---|
| 1 | Breast cancer | 0.966 | 0.966 |
| 2 | Ionosphere | 0.709 | 0.704 |
| 3 | Primary tumour | 0.508 | 0.302 |
| 4 | Soybean small | 1.000 | 1.000 |
| 5 | Zoo | 0.581 | 0.584 |
| 6 | Heart disease case 2 | 0.802 | 0.802 |

## CONCLUSION

Several investigations to avoid empty groups in the k-medoids algorithm have developed. This paper proposed another view method that guarantees there are no empty groups and identical objects as non-unique medoids in the same cluster, either in the initial groups or final groups. The k-medoids algorithms based on distance variance blocks (proposed method) can handle clustering datasets containing identical objects as non-unique medoids. The new algorithm applies to any type of data, i.e. categorical, numerical or mixed data. Evaluation of new methods on the six real datasets, i.e. breast cancer, ionosphere, primary tumour, soybean small, zoo and heart disease case 2, produce the adjusted Rand index; and clustering accuracy is comparable with another method. The Var-KM algorithm (new method) enriches the reference on partitioned-based clustering of data sets.

## REFERENCES

Behzadi, S., Muller, N.S., Plant, C., and Bohm, C., (2020), *Clustering of mixed-type data considering concept hierarchies: problem specification and algorithm*, International Journal of Data Science and Analysics, 10: 233-248.

Budiaji, W., Leisch, F., (2019), Simple k-medoids partitioning algorithm for mixed variable data, in Algorithms, 12(177): 1-15.

Dinata, R.K., Retno, S., and Hasdyna, N. (2021). Minimization of the Numer of Iterations in K-Medoids Clustering with Purity Algorithm. Revue d'Intelligence Articielle, 35(3): 193-199.

Everitt, B.S., Landau, S. Leese, M., and Stahl, D. (2011). Cluster analysis, 5th edn., John Wiley & Sons., Ltd., Publication. ISBN: 978-0-470-97844-3. pp. 49-50.

Jajuga, K. (2000). Standardization of Data Set under Different Measurement Scales.

Ji., J., Li, R., Pang, W., He, F., Feng, G., Zhao, X. (2021). A Multi-View Clustering Algorithm for Mixed Numeric and Categorical Data. IEEE Access, 10: 24913-24924.

Ji., J., Pang, W., Li, Z., He, F., Feng, G., Zhao, X., (2020). Clustering mixed numeric and categorical data with cuckoo search. IEEE Access, 8: 30988-31003.

Johnson, R.A., Wichern, D.W. (2009). Applied Multivariate Statistical Analysis, John Wiley & Sons. ISBN: 0-13-187715-1. pp. 680-695.

Kariyam, Abdurakhman, Subanar, and Herni, U. (2022). The Initialization of Flexible K-Medoids Partitioning Methods Using a Combination of Deviation and Sum of Variable Values. Mathematics and Statistics, 10(5): 895-908.

Kaufman, L., Rousseeuw, P.J. (1990). Finding groups in data: An introduction to cluster analysis, John Wiley & Sons, Inc., New YorkISBN: 0-471-73578-7. pp. 28-38, 68-71, 102-104.

Lichman, M. (2021-2022). UCI Machine Learning Repository, University of California: Irvine, CA, USA, accessed on Nov. 10, 2021, and Jun. 15, 2022)

Nitesh, S., Chawda, B., Vasant, A. (2022). An improved K-medoids clustering approach based on the crow search algorithm. Journal of Computational Mathematics and Data Science. 3: 100034. https://doi.org/10.016/j.jcmds.2022.100034

Oesting, M., and Schnurr, A., (2020), *Ordinal patterns in clusters of subsequent extremes of regularly varying time series*, Applied Intelligence, Vol. 50, p.1498-1509

Park, H.S., Jun, C.H. (2009). A Simple and Fast Algorithm for K-Medoids Clustering. Expert System with Applications, 36(2): 3336–3341.

Schubert, E., and Rousseeuw, P.J. (2021). Fast and eager k-medoids clustering: O(k) runtime improvement of the PAM, CLARA, and CLARANS algorithms. Information Systems, Elsevier, 101(2021): 101804.

Selosse, M., Jacques, J., and Bienachki, C., (2019), *Model-based co-clustering for type data*, manuscript version under the Elsevier user Licence.

Warrens, M.J., van der Hoef, H. (2022). Understanding the Adjusted Rand Index and Other Partition Comparison Indices Based on Counting Object Pairs. *Journal of Classification.* https://doi.org/10.1007/s00357-022-09413-z

Wu, J. (2012). Advance in K-means Clustering: A Data Mining Thinking. Springer-Verlag, Berlin Heidelberg.

Xu, R., and Wunsch, D.C.II. (2009). Clustering. John Wiley & Sons, Inc., Hoboken, New Jersey. ISBN: 978-0-470-27680-8. pp. 23-24.

Yu, D., Liu, G., Guo, M., Liu, X. (2018). An improved k-medoids algorithm based on step increasing and optimizing medoids. Expert System With Applications, 92(2018): 464-473.

Yuan, F., Yang, Y., Yuan, T. (2020). A dissimilarity measure for mixed nominal and ordinal attribute data in k-Modes algorithm. Applied Intelligence, 50: 1498-1509,

Zorn, C., (2003). Agglomerative Clustering of Rankings Data, with Application to Prison Rodeo Events. Department of Political Science, Emory Universit, Atlanta, GA 30322.