Building of Informatics, Technology and Science (BITS)
Volume 4, No 4, Maret 2023 Page: 1705−1713
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i4.2791

Optimization in Time and Score using IID Algorithm for K-Modes Clustering

Farah Yulianti[1,*], Tjong Wan Sen[2]

Faculty of Computing, President University, Bekasi, Indonesia
Email: [1,*]reviewfarmei20@gmail.com, [2]wansen@president.ac.id
Correspondence Author Email: reviewfarmei20@gmail.com
Submitted: **29/12/2022**; Accepted: **27/03/2023**; Published**: 31/03/2023**

Abstract−Nowadays, there are numerous methods for analyzing data, one of which is cluster analysis. Because most practical data in today's analysis contains categorical attributes, categorical data clustering has recently received a lot of attention. To cluster categorical data, unsupervised machine learning techniques, which used frequency-based methods, such as K-Mode's clustering are used. The K-Modes algorithm takes advantage of the differences between the data points (total mismatches or dissimilarities). The lower the dissimilarities, the more similar the data points, and thus the better the cluster. This paper aims to improve K-Mode's clustering performance by incorporating the intercluster and intracluster dissimilarity measure, or IID measure, into the K-Modes algorithm rather than just using the standard simple-matching method to increase the algorithm's accuracy and execution time. This combined algorithm improves the accuracy and execution time of the K-Modes algorithm. As a result, this algorithm can be used as an alternative to better cluster categorical data.

Keywords: Clustering Algorithm; K-Modes; InterIntra-Cluster; Dissimilarities

1. INTRODUCTION

Machine learning is continuously becoming a hot topic in the world. Machine learning has permeated all aspects of life. Starting with straightforward tasks like setting up a system to learn when to turn on or off the lights, predicting whether it will rain or not, and even predicting when an engine will break down. Complex examples include categorizing films into different genres, determining whether humans wear masks or not, and learning how to predict the outcome of cryptocurrencies and stocks. [1]

Machine learning is separated into two types which are unsupervised and supervised learning. Supervised learning is relying on training data to be made into a model for the input data or testing data while unsupervised learning relies more on whole data to be grouped into similar characteristic data [2]. Unsupervised learning is commonly used for data without any labels or scores for the training data. One of the algorithms used for unsupervised learning is the K-Modes algorithm. K-Modes are created as an adjusted version of the K-Means algorithm to cluster categorical data [3]–[5]. K-Modes is a clustering method for categorical data that has been widely used in a variety of applications, including market segmentation and image classification [12]. To determine the distance between two objects, it employs a simple matching dissimilarity measure known as the Hamming distance rather than the Euclidean distance. Furthermore, it represents the cluster centroids using modes rather than means.

Some issues have an impact on the quality of clusters created with the K-Modes algorithm. First, is the cluster assignment selection. [5] and then the measures of both similarity and dissimilarity of the non-numerical data [4], which sometimes cause a very slow execution time and mediocre accuracy. Many improvements have been made to the K-Modes algorithm to improve either the accuracy or the execution time. Although it has been improved numerous times, it is believed that there are still research gaps in the K-Modes clustering algorithm that could be improved in terms of accuracy and/or algorithm execution time. [6]

The K-Modes algorithm was first proposed in 1998 by Huang to tackle the problem of clustering categorical data. The algorithm has had a few upgrades over the years with the latest one from Huang in 2009 in terms of initialization [3]. Then another improvement was proposed for the time complexity and accuracy by using a density-based approach for the initialization Cao in 2012 [4,13]. The study's novelty lies in the dissimilarity measure of the K-Modes clustering by changing it from just using a simple matching method as proposed by Huang [3] into using a method known as intercluster and intracluster dissimilarity measure or will be mentioned as IID (InterIntra-Cluster Dissimilarity) measure. This method combines the object's relationship with the different clusters or will be called Inter-cluster and with the relationship inside the same cluster Intra-cluster. The IID method was first implemented in the K-Means algorithm as an added method for feature selection in the algorithm [7] and then also implemented in other clustering algorithms such as TaxMap [14], heuristic [15], and Mobile Sink algorithm [16,17]. The IID implementation on those algorithms has numerous good results such as improved accuracy and execution time.

2. RESEARCH METHODOLOGY

The design of the research is stated below:

2.1 Research Object

Building of Informatics, Technology and Science (BITS)
Volume 4, No 4, Maret 2023 Page: 1705−1713
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i4.2791

The K-Modes clustering algorithm is the focus of the thesis. To improve K-Modes, the distance measure is changed rather than using simple matching, and the result is compared to the standard K-Modes algorithm or regular K-Means. [8]

2.2 Data Collection

This section contains the explanation, the structure, and the content of the data used in this thesis. The datasets used are taken from the University of California, Irvine (UCI) Machine Learning repository.

2.3 Proposed Algorithm and Implementation

This section contains an explanation of the proposed algorithm to enhance K-Modes called IID (InterIntra Cluster Dissimilarity) which combines the object's relationship with all clusters (inter-cluster) with that within clusters (intra-cluster). The algorithm will be used to change the dissimilarity measure within K- The modes algorithm to refine the accuracy of the algorithm hypothetically [7]. The research is illustrated below in Figure 2 :

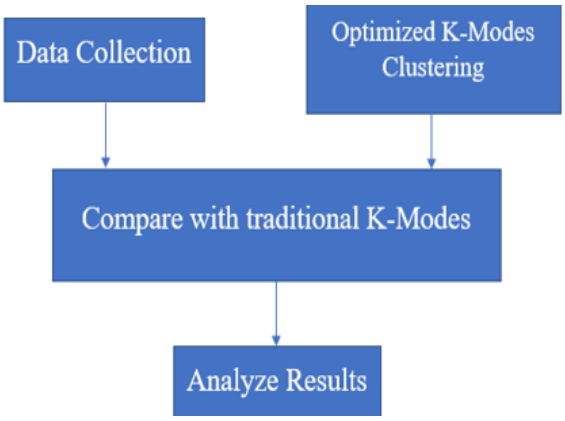

Figure 1. Process of K-Modes Optimization

As shown in Figure 1, after data collection is finished, then comparation between the traditional K-Modes and Optimized K-Modes will be done by looking at the time and accuracy of those algorithms. Then the result will be analyzed to determine whether or not optimization using IID algorithm on K-Modes is capable on increasing both the accuracy and time.

2.4 Data Collection

The K-Modes with the IID algorithm are tested using the datasets obtained from the UCI Machine Learning repository. The Chess, Nursery, Wisconsin Breast Cancer, and Lung Cancer datasets were used. The information about the datasets can be seen in Table 1.

Table 1. Datasets

No	Dataset	Attribute characteristics	# of objects	# of attributes	# of Class
1	Chess	Categorical	3196	36	2
2	Nursery	Categorical	12960	8	5
3	Wisconsin Breast Cancer	Numerical	699	10	2
4	Lung Cancer	Numerical	32	56	3

In Table 1, there are 4 datasets used on this paper. The first dataset is using Chess dataset which contains categorical variables with 3196 number of objects, 36 number of attributes, and two number of classes. The second dataset is using Nursery dataset which contains categorical variables with 12960 number of objects, eight number of attributes, and five number of classes. The third dataset is using Wisconsin Breast Cancer dataset which contains numerical variables with 699 number of objects, ten number of attributes, and two number of classes. The fourth dataset is using Lung Cancer dataset which contains numerical variables with 32 number of objects, 56 number of attributes, and three number of classes.

2.5 Proposed Algorithm and Implementation

The proposed method is called the InterIntra-Cluster Dissimilarity measure or will be known as IID would be applied in the distance measure of the K-Modes algorithm. IID considers the relationship between the object and all clusters as well as that within clusters instead of just simple matching. The Pseudocode of this algorithm can be reviewed below in Figure 3:

Building of Informatics, Technology and Science (BITS)
Volume 4, No 4, Maret 2023 Page: 1705−1713
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i4.2791

Algorithm 1: Pseudocode of K-Modes with IID

```
Input  : Dataset (U), initial cluster number (k)
Output: Clusters
Randomly choose k distinct objects as initial modes from U;
for l = 0 to k do
    for i = 1 to n do
        Calculating Dissimilarity K-Modes IID according to Eq 1 − 5;
        Calculating sumDissimilarity according to
         K-Modes Huang Algorithm;
        newDissimilarity = sumDissimilarity;
    end
end
if Dissimilarity K-Modes IID ≤ DissimilarityK − Modes then
    Classify ith object to lth cluster;
end
while newDissimilarity ≠ oldDissimilarity do
    oldDissimilarity = newDissimilarity;
    Update modes according to the K-Modes algorithm;
    for l = 0 to k do
        for i = 1 to n do
            Calculating Dissimilarity K-Modes IID according to Eq 1 −
             5;
            Calculating sumDissimilarity according to
             K-Modes Huang Algorithm;
            newDissimilarity = sumDissimilarity;
        end
    end
end
```

Figure 2. Pseudocode of K-Modes with IID Algorithm

A shown in Figure 2, this proposed approach could minimize the K-Modes cluster problem because it replaces the initial simple matching from the K-Modes algorithm with the InterIntra Cluster dissimilarity (IID) measure to be able to increase the accuracy and execution time of the algorithm by assigning supposed similar data points to more appropriate cluster rather than just assigning it into a random cluster.

3. RESULT AND DISCUSSION

The proposed algorithm's performance in terms of accuracy and execution time is compared to standard K-Modes with Huang [3] and Cao [5] initialization. Because both of the standard K-Modes are executed with a known class in their cluster parameter, the standard K-Modes is executed several times in the experiment to obtain the average value of the result [9], [10].

To evaluate the clustering result, several parameters are passed into the K-Modes model provided by Python Package Index, or PyPI. The experiments are carried out on a Laptop with an AMD Quad-Core FX-9800P processor and 8GB memory running the Windows 10 operating system. All algorithms are written in Python on Jupyter Notebook.

To evaluate the performance of the algorithm, we use two kinds of clustering evaluations, which are the Fowlkes-Mallows score and RandIndex. Let Table 2 be an example for Fowlkes-Mallows and RandIndex as follows:

Table 2. Score Labels

Labels		Cluster Results	
		0	**1**
Measured Data	**0**	1	1
	1	1	0

As shown in table 2 above, the True Positive (TP) of the Fowlkes-Mallows and RandIndex score is the number of data points that are present in both Cluster 1 (C1) and Cluster 2 (C2). The False Positive (FP) of the Fowlkes-Mallows score is the number of data points that are present only in Cluster 1 (C1) while the False Negative (FN) of the Fowlkes-Mallows and RandIndex score is the number of data points that are only present in Cluster 2 (C2). The equation for Fowlkes-Mallows score as Equation 6 is shown below:

$$FM = \frac{TP}{\sqrt{(TP+FP).(TP+FN)}} \quad (1)$$

Building of Informatics, Technology and Science (BITS)
Volume 4, No 4, Maret 2023 Page: 1705−1713
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i4.2791

The equation for the RandIndex score as Equation 7 is shown below:

$$RI = \frac{TP+TN}{TP+FP+FN+TN} \quad (2)$$

The adjusted RandIndex formula as shown in equation 7 above is a chance-corrected version of the RandIndex. A benchmark is established by utilizing a few comparisons between each cluster which is indicated by a random clustering model's expected similarity. While the RandIndex can only return values between 0 and +1, the adjusted RandIndex can return negative values if the index is less than the expected index [11]. The equation of Adjusted RandIndex as Equation 8 and Equation 9 are shown below:

$$ARI = \frac{TP-Expected\ RI}{Max(RI)-Expected\ RI} \quad (3)$$

$$Max(RI) = \frac{FP+FN+2TP}{2} \quad (4)$$

In the experiments, we compare the K-Modes with the IID algorithm with the K-Modes with Huang and Cao initialization. The three algorithms are run sequentially on all the datasets. The parameters that are passed to the K-Modes algorithm are as follows: The cluster number is the number of clusters that must be formed and generated, and the cluster number will be set based on the classes of each dataset respectively. The maximum iteration is set to 100 to limit the iteration and the number of initiation will be set to 25 for all datasets which is the best run for each of the datasets. The Repeat parameter is used to repeat the process a set number of times.

The Fowlkes-Mallows and RandIndex score results with the execution time of the three algorithms on each dataset can be shown in Table 3-6. It shows that the K-Modes with the IID algorithm get the highest score out of all the algorithms in each repetition.

Table 3. Performance Evaluation On Chess Dataset

Repeat	Algorithm	Fowlkes-Mallows (%)	Adjusted RandIndex	Execution Time (s)
1	K-Modes Huang	55.2	0.031	88
	K-Modes Cao	58.3	0.016	**7**
	K-Modes IID	**61.1**	**0.095**	58
2	K-Modes Huang	55.2	0.031	194
	K-Modes Cao	58.3	0.016	**18**
	K-Modes IID	**59.3**	**0.067**	91
3	K-Modes Huang	54.8	0.015	143
	K-Modes Cao	58.3	0.016	**9**
	K-Modes IID	**59.3**	**0.062**	84
4	K-Modes Huang	55.3	0.022	90
	K-Modes Cao	58.3	0.016	**9**
	K-Modes IID	**61.9**	**0.1**	67
5	K-Modes Huang	55.3	0.022	106
	K-Modes Cao	58.3	0.016	**8**
	K-Modes IID	**62.4**	**0.055**	59

The Fowlkes-Mallows score of the Chess dataset is illustrated below in Figure 4 :

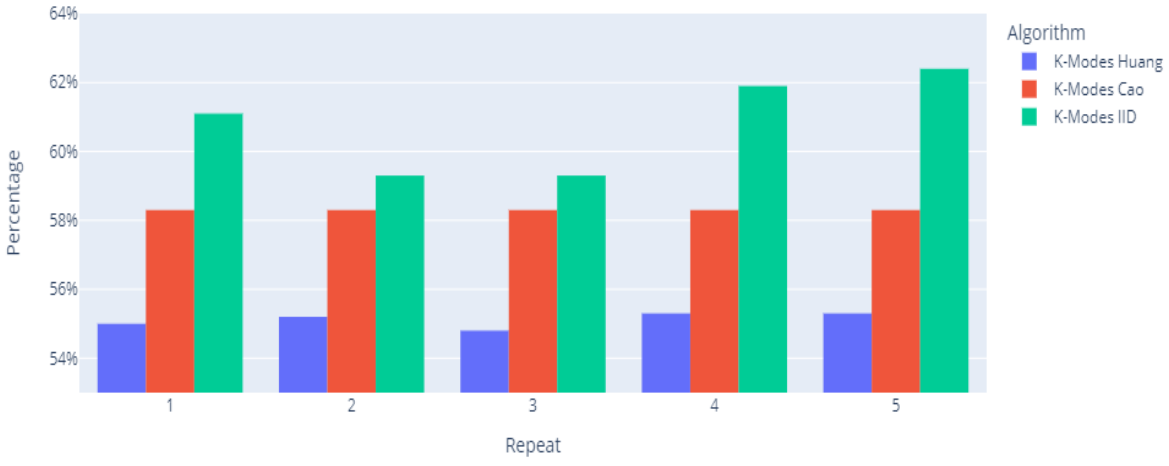

Figure 4. Fowlkes-Mallows Score of Chess Dataset

The RandIndex score of the Chess dataset is illustrated below in Figure 5 :

Building of Informatics, Technology and Science (BITS)
Volume 4, No 4, Maret 2023 Page: 1705−1713
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i4.2791

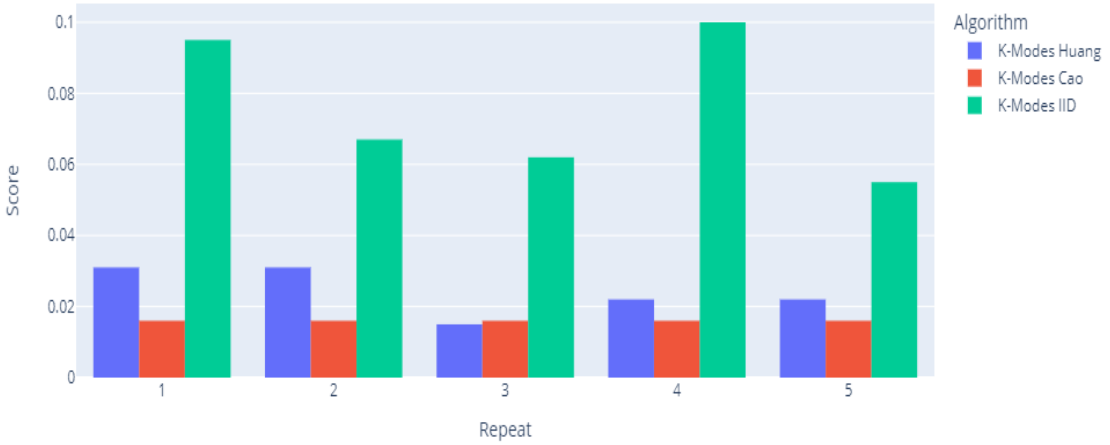

Figure 5. Adjusted RandIndex Score of Chess Dataset

As shown in table 3 above, the result score of Fowlkes-Mallow and Adjusted RandIndex on the chess dataset can be inferred that the K-Modes with IID is higher than both K-Modes Huang and Cao. Although, it still lost in terms of time spent by K-Modes with Cao algorithm. The illustrated result of K-Modes with IID performacne can be seen on Figure 4 for Fowlkes-Mallow result and Figure 5 for Adjusted RandIndex score.

Table 4. Performance Evaluation On Nursery Dataset

Repeat	Algorithm	Fowlkes-Mallows (%)	Adjusted RandIndex	Execution Time (s)
	K-Modes Huang	27.8	0.029	148
1	K-Modes Cao	31.5	0.064	8
	K-Modes IID	35.8	0.127	99
	K-Modes Huang	31.3	0.062	161
2	K-Modes Cao	31.5	0.064	8
	K-Modes IID	32.9	0.086	132
	K-Modes Huang	29.4	0.039	160
3	K-Modes Cao	31.5	0.064	8
	K-Modes IID	32.7	0.083	133
	K-Modes Huang	30	0.051	161
4	K-Modes Cao	31.5	0.064	8
	K-Modes IID	33.3	0.092	152
	K-Modes Huang	30.1	0.045	177
5	K-Modes Cao	31.5	0.064	8
	K-Modes IID	35.3	0.120	161

The Fowlkes-Mallows score of the Nursery dataset is illustrated below in Figure 6 :

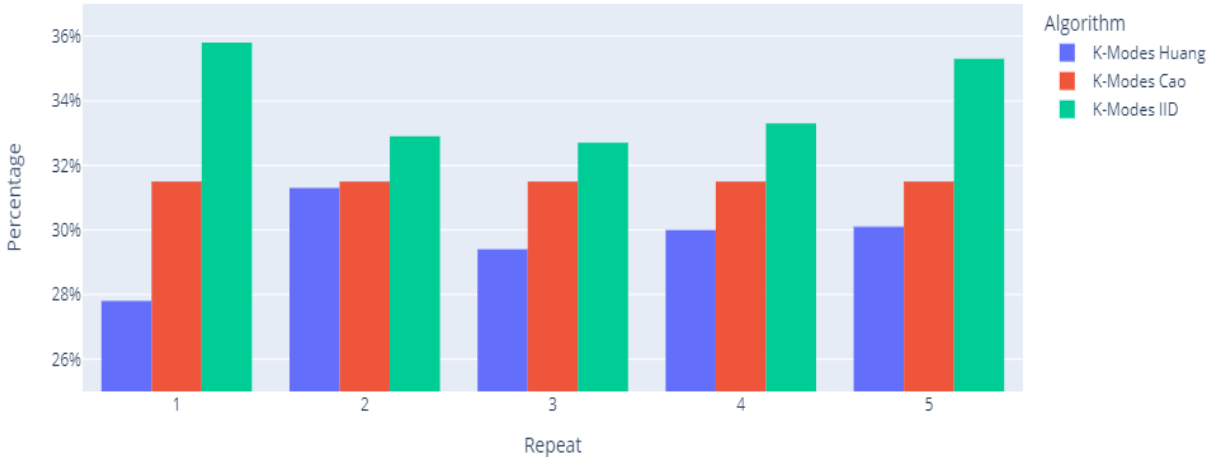

Figure 6. Fowlkes-Mallows Score of Nursery Dataset

The RandIndex score of the Nursery dataset is illustrated below in Figure 7 :

Building of Informatics, Technology and Science (BITS)
Volume 4, No 4, Maret 2023 Page: 1705−1713
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i4.2791

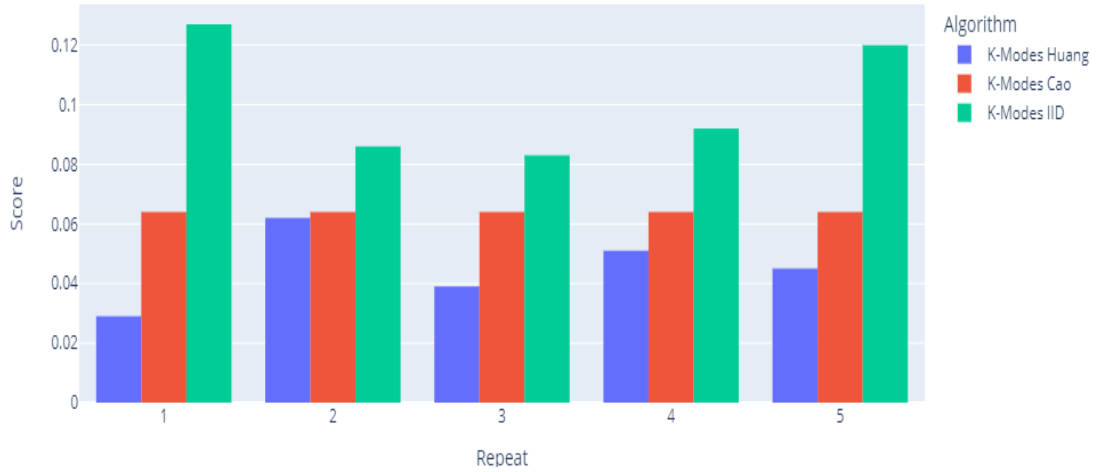

Figure 7. RandIndex Score of Nursery Dataset

As shown in table 4 above, the result score of Fowlkes-Mallow and Adjusted RandIndex on the nursery dataset can be inferred that the K-Modes with IID is higher than both K-Modes Huang and Cao. Although, it still lost in terms of time spent by K-Modes with Cao algorithm. The illustrated result of K-Modes with IID performacne can be seen on Figure 6 for Fowlkes-Mallow result and Figure 7 for Adjusted RandIndex score

Table 5. Performance Evaluation On Wisconsin Breast Cancer Dataset

Repeat	Algorithm	Fowlkes-Mallows (%)	Adjusted RandIndex	Execution Time (s)
	K-Modes Huang	86	0.669	8
1	K-Modes Cao	86	0.669	0.59
	K-Modes IID	90.9	0.796	1.4
	K-Modes Huang	86	0.669	8
2	K-Modes Cao	86	0.669	0.59
	K-Modes IID	90.9	0.796	1.53
	K-Modes Huang	86	0.669	6.63
3	K-Modes Cao	86	0.669	0.52
	K-Modes IID	90.9	0.796	1.53
	K-Modes Huang	86	0.669	9.19
4	K-Modes Cao	86	0.669	0.57
	K-Modes IID	90.9	0.796	1.6
	K-Modes Huang	86	0.669	6.55
5	K-Modes Cao	86	0.669	0.51
	K-Modes IID	90.9	0.796	1.62

The Fowlkes-Mallows score of the WBC dataset is illustrated below in Figure 8 :

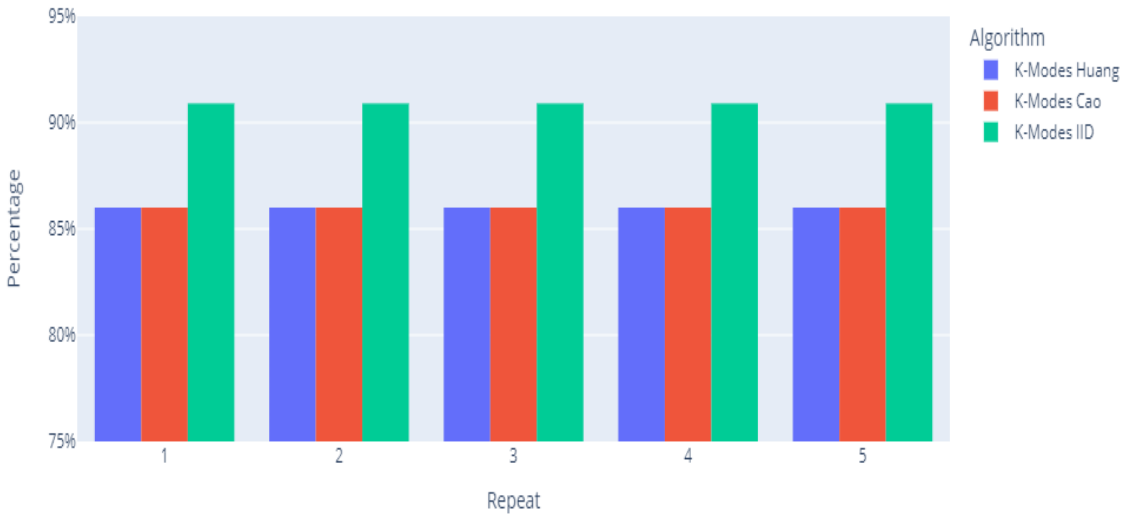

Figure 8. Fowlkes-Mallows Score of WBC Dataset

The RandIndex score of the WBC dataset is illustrated below in Figure 9 :

Building of Informatics, Technology and Science (BITS)
Volume 4, No 4, Maret 2023 Page: 1705−1713
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i4.2791

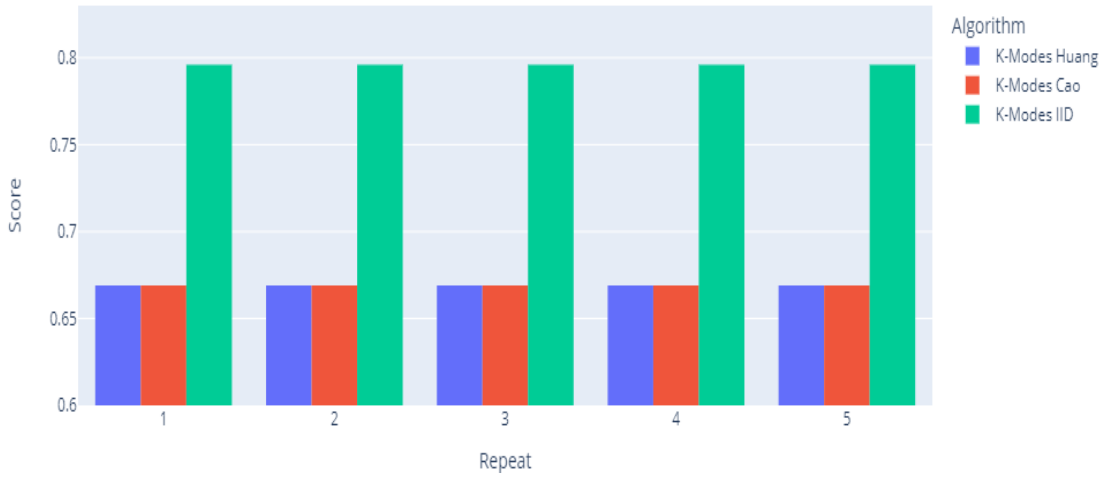

Figure 9. RandIndex Score of WBC Dataset

As shown in table 5 above, the result score of Fowlkes-Mallow and Adjusted RandIndex on the WBC dataset can be inferred that the K-Modes with IID is higher than both K-Modes Huang and Cao. Although, it still lost in terms of time spent by K-Modes with Cao algorithm. The illustrated result of K-Modes with IID performacne can be seen on Figure 8 for Fowlkes-Mallow result and Figure 9 for Adjusted RandIndex score

Table 6. Performance Lung Cancer Dataset

Repeat	Algorithm	Fowlkes-Mallows (%)	Adjusted RandIndex	Execution Time (s)
	K-Modes Huang	53.1	0.117	2.96
1	K-Modes Cao	46.8	-0.01	0.13
	K-Modes IID	60.9	0.2	2
	K-Modes Huang	53.3	0.127	3.72
2	K-Modes Cao	46.8	-0.01	0.2
	K-Modes IID	63.4	0.266	2
	K-Modes Huang	50.2	0.084	3.72
3	K-Modes Cao	46.8	-0.01	0.24
	K-Modes IID	59.7	0.204	1.97
	K-Modes Huang	49.9	0.087	2.96
4	K-Modes Cao	46.8	-0.01	0.17
	K-Modes IID	60.6	0.228	2.47
	K-Modes Huang	55.1	0.169	2.99
5	K-Modes Cao	46.8	-0.01	0.14
	K-Modes IID	69.5	0.314	2.17

The Fowlkes-Mallows score of the Lung Cancer dataset is illustrated below in Figure 10 :

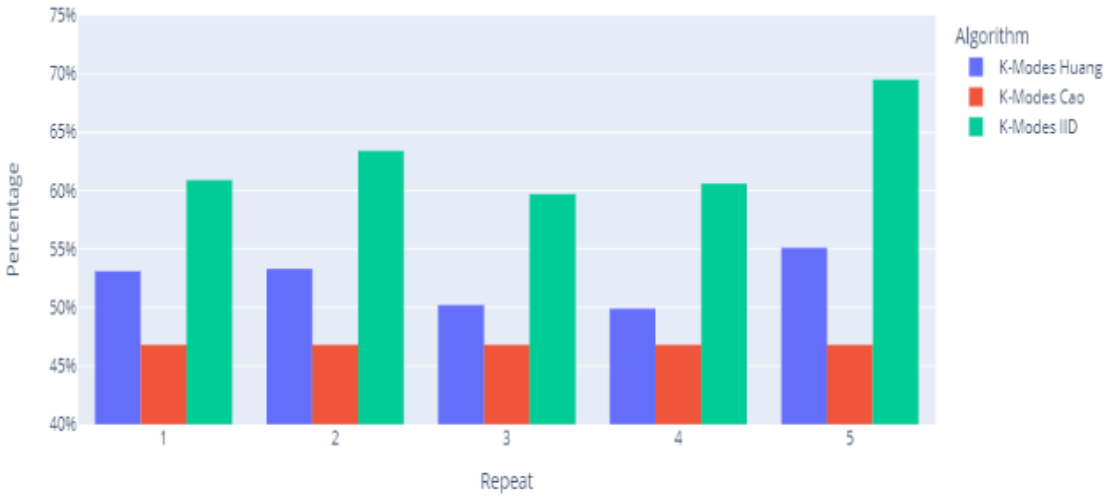

Figure 10. Fowlkes-Mallows Score of Lung Cancer Dataset

The RandIndex score of the Lung dataset is illustrated below in Figure 11 :

Building of Informatics, Technology and Science (BITS)
Volume 4, No 4, Maret 2023 Page: 1705−1713
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i4.2791

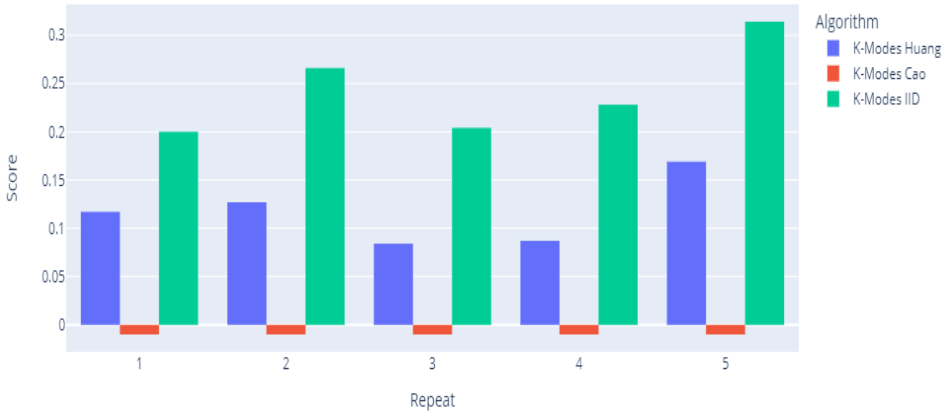

Figure 11. RandIndex Score of Lung Cancer Dataset

As shown in table 6 above, the result score of Fowlkes-Mallow and Adjusted RandIndex on the lung cancer dataset can be inferred that the K-Modes with IID is higher than both K-Modes Huang and Cao. Although, it still lost in terms of time spent by K-Modes with Cao algorithm. The illustrated result of K-Modes with IID performacne can be seen on Figure 10 for Fowlkes-Mallow result and Figure 11 for Adjusted RandIndex score

Based on the results of the experiment above in table 3-6 and figure 4-11, the K-Modes with the IID algorithm are stable in terms of accuracy, meaning have a significant improvement, and can be used as an alternative to better cluster categorical data. Although execution time still falls behind K-Modes with Cao initialization, it is still faster and more reliable rather than K-Modes with Huang initialization.

4. CONCLUSION

This study examines the use of Intercluster and Intracluster dissimilarity measures of the K-Modes algorithm for categorical data. Based on this, we propose to combine K-mode with the IID algorithm to improve categorical data grouping. This measure is used to improve the performance of existing K-Mode algorithms. On four real data sets from the UCI, we tested K-Modes with the IID algorithm. Several conclusions can be drawn from this experiment namely that the accuracy and execution time of K-Mode is improved by combining the K-Modes algorithm with the InterIntra Dissimilarity or IID measure, K-Mode with IID improves K-Mode accuracy with Huang and Cao initialization significantly, K-Mode with IID outperforms K-Mode with Huang initialization in terms of execution time but still lags behind K-Mode with Cao initialization and K-Mode with IID has better accuracy based on Fowlkes-Mallows and RandIndex Score. In the future, we hope to improve the algorithm by adding some improvements, especially in terms of making the execution time faster.

REFERENCES

[1] D.-T. Dinh and V.-N. Huynh, “k-PbC: an improved cluster center initialization for categorical data clustering,” Applied Intelligence, vol. 50, no. 8, pp. 2610–2632, 2020.
[2] Pal, S. K., & Pal, M. A Comparative Study of Initialization Methods for K-Means-Type Clustering Algorithms IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021.
[3] Kuo, R. J., & Nguyen, T. P. Q. Genetic intuitionistic weighted fuzzy k-modes algorithm for categorical data. Neurocomputing, 330, 116-126, 2019.
[4] Zafar, A., & Swarupa Rani, K. Novel Initialization Strategy for K-modes Clustering Algorithm. In Proceedings of International Conference on Big Data, Machine Learning and Applications (pp. 89-100). Springer, Singapore, 2021.
[5] F. Cao et al., “An algorithm for clustering categorical data with set-valued features,” IEEE Trans Neural Netw Learn Syst, vol. 29, no. 10, pp. 4593–4606, 2017.
[6] Xiao, Y., Huang, C., Huang, J., Kaku, I., & Xu, Y. Optimal mathematical programming and variable neighborhood search for k-modes categorical data clustering. Pattern Recognition, 90, 183-195, 2019.
[7] Wang, Y., & Zhang, Y. A K-Means Clustering-Based Hybrid Offspring Generation Mechanism in Evolutionary Multi-Objective Optimization. IEEE Access, 9, 167642-167651, 2021.
[8] Guo, J., Li, X., Li, X., & Li, Y. “Gaussian Mixture Model for Mixed Data Types”. IEEE Transactions on Cybernetics, 2021.
[9] Oskouei, A. G., Balafar, M. A., & Motamed, C. FKMAWCW: categorical fuzzy k-modes clustering with automated attribute-weight and cluster-weight learning. Chaos, Solitons & Fractals, 153, 111494, 2021.
[10] Y. Zhang, Y. Yang, T. Li, and H. Fujita, “A multitask multiview clustering algorithm in heterogeneous situations based on LLE and LE,” Knowl Based Syst, vol. 163, pp. 776–786, 2019.
[11] A. J. Gates and Y.-Y. Ahn, “The impact of random models on clustering similarity,” arXiv preprint arXiv:1701.06508, 2017.
[12] Everitt, B. S., Landau, S., & Leese, M. Handbook of cluster analysis. CRC press, 2019.
[13] Yuan, F., Yang, Y., & Yuan, T. A dissimilarity measure for mixed nominal and ordinal attribute data in k-Modes algorithm. Applied Intelligence, 50(5), 1498-1509, 2020.
[14] Alves, G., Couceiro, M., & Napoli, A. Similarity Measure Selection for Categorical Data Clustering, 2019.

Building of Informatics, Technology and Science (BITS)
Volume 4, No 4, Maret 2023 Page: 1705−1713
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i4.2791

[15] Jahwar, A. F., & Abdulazeez, A. M. Meta-heuristic algorithms for K-means clustering: A review. PalArch's Journal of Archaeology of Egypt/Egyptology, 17(7), 12002-12020, 2020.
[16] Gharaei, N., Bakar, K. A., Hashim, S. Z. M., & Pourasl, A. H. Inter-and intra-cluster movement of mobile sink algorithms for cluster-based networks to enhance the network lifetime. Ad Hoc Networks, 85, 60-70, 2019.
[17] Wei, Q., Bai, K., Zhou, L., Hu, Z., Jin, Y., & Li, J. A cluster-based energy optimization algorithm in wireless sensor networks with mobile sink. Sensors, 21(7), 2523, 2021.