A Simplified Evaluation Method Used for Unsupervised Learning Algorithms

Ying Bai[1,*] and Dali Wang[2]

[1]Department of Computer Science and Engineering, Johnson C. Smith University, Charlotte, USA
[2]Department of Physics and Computer Science, Christopher Newport University, Newport News, USA
Email: ybai@jcsu.edu (Y.B.); dwang@cnu.edu (D.W.)
[*]Corresponding author

***Abstract*—Professional evaluations for unsupervised learning algorithms are a crystal important and challenging topic due to its input-output specialty. There are no output labels available in unsupervised learning models to work as an evaluation criterion to check and compare the predicted outputs with the actual outputs. In other words, it is relatively difficult to effectively estimate or validate the correctness or efficiency of any unsupervised learning algorithms with regular methods used in supervised learning studies.**

In this study, we developed a simplified evaluation method by using average or mean values for testing observations and comparing them with the mean values of clustered observations to identify their correctness of their clustering. This method can be used to effectively perform some evaluations for unsupervised learning models. A comparison between our method with some traditional evaluation methods, such as Silhouette Scores, is performed and the comparison result shows that our method is better than that traditional method.

The experimental study confirmed the effectiveness of our evaluation methods used for unsupervised learning algorithms, and the correct evaluation rate is up to 97%.

***Keywords*—Evaluations for unsupervised learning algorithm, internal validation metrics, Silhouette scores, Calinski-Harabasz index, Davies-Bouldin index, Dunn index, Diabetes dataset.**

I. INTRODUCTION

Unsupervised learning (UL) is one of the most popular study fields in machine learning (ML), and it plays important roles in AI implementations and applications in recent years. However, for most UL algorithms and models, a key and challenging issue is how to correctly and accurately evaluate and validate a UL algorithm to make sure that it is a good candidate in most AI implementations. Although UL is a popular algorithm and widely implemented in all aspects in our world, few of studies related to evaluations of UL with new methods has been reported so far. Most of the research is related to using traditional validation methods, either internal or external validation methods, to evaluate various UL models.

[1] provides some extensions of cross-validation methods to evaluate UL algorithms. Julio-Omar et al. provided a full review for internal and external validation methods [2]. Andreas Vlachos suggested using an in-context evaluation to evaluate unsupervised learning models [3]. Garg, Vikas, and Adam T. Kalai introduced a framework to transfer knowledge acquired from a repository of (heterogeneous) supervised datasets to new unsupervised datasets [4]. Martin et al. provided a survey for most internal evaluation strategies used for unsupervised learning but concluded that none of the existing and adapted strategies would be practically useful for evaluation of unsupervised learning models [5]. Portela et al. also provided a survey for most evaluation methods used in supervised and unsupervised learning models, such as SVM and K-Means, but concluded that not all evaluation methods would be good for unsupervised learning models [6]. Padmavathi et al. evaluated six unsupervised learning models with internal and external validation metrics with a conclusion that the K-Means method has the highest clustering accuracy of 88% for mobile malware detection [7]. Khoei et al. evaluated the performance of four unsupervised deep learning models with traditional methods, such as Calinski harabasz score, Silhouette Score, homogeneity, completeness, and V-measure, with a conclusion that the Convolutional Auto-Encoder has the best performance results among the other unsupervised deep learning models [8]. O. Toyin et al. evaluated three popular unsupervised learning algorithms, KNN, K-Mean and Hierarchical Clustering (HC), with locally sourced data from Ekiti State Teaching Hospital, Ado Ekiti, and appraised using standard metrics [9].

R. Barreto, L. Gomes and Z. Valeused used a set of unsupervised learning evaluation metrics to evaluate members' data for the planned demand response events [10]. A. U. Haq et al. used *k* fold cross validation method to evaluate their unsupervised learning models to find optimal model used to detect breast cancers [11]. Zou et al. proposed a pseudo-label generation and ranking guidance (PGRG) framework for label-free PIQA to transfer an unsupervised learning model to a supervised learning model to improve the palmprint recognition processing [12].

For all the above methods used to evaluate UL models, they are either very complicated in principles and implementations or just traditional validation metrics with limited validation accuracy.

In this study, we proposed a simplified evaluation method by using average or mean values of testing observations and comparing them with the mean values of clustered observations to identify their correctness of their clustering. This method can be used to effectively perform some evaluations for unsupervised learning models. A comparison between our method with some traditional evaluation methods, such as Silhouette Scores, is performed and the comparison result shows that our method is better than that of traditional method.

This study is divided into five sections. After the introduction part, a quick review for some traditional validation metrics is given in section 2. Detailed introduction

https://doi.org/10.34630/cip.vi.7656

for our simplified evaluation method is provided in part 3. Section 4 shows a comparison of evaluation results between our simplified evaluation method and traditional validation method, Silhouette score, to illustrate the effectiveness of our method.

II. Traditional Evaluation Methods for UL

Traditionally, the evaluation methods used for unsupervised learning models include Internal Validation Metrics, such as Silhouette Score, Calinski-Harabasz Index, Davies-Bouldin Index, Dunn Index, and Inertia, as well as External Validation Metrics, such as Adjusted Rand Index, Mutual Information Score and Cophenetic Correlation Coefficient [13].

Relatively speaking, the internal validation metrics are more popular due to their simplicity and easy to use. In this study, we will concentrate on this kind of method with more discussions about them.

Silhouette Score measures how well an observation fits into its assigned cluster compared to other clusters. Calinski-Harabasz Index, also known as the Variance Ratio Criterion, evaluates the cluster validity based on the average between-cluster and within-cluster sum of squares. The principle of these two methods is basically based on the distance ratios between the same cluster and different clusters. Due to their similarity, we use the silhouette score to illustrate the function of these two methods.

The silhouette score is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation) [14]. The Silhouette score is exactly a value between -1 and 1, where higher values indicate better clustering. This index is especially useful for high-dimensional datasets used for clustering purposes.

Suppose a collection of input data has been clustered by using any clustering technique, such as k-means. For a data point $i \in$ Ci (data point i in the cluster Ci, $i = 1 \sim$ n), let a(i) be the mean distance between i and all other data points in the same cluster, where n is the number of points belonging to cluster i,

$$a(i) = \frac{1}{n-1}\sum_{j \in C_i, i \neq j} d(i,j) \tag{2.1}$$

Then we define the mean dissimilarity of point i to all other cluster Cj as the mean of the distance from i to all points in Cj as b(i) as

$$b(i) = \min_{k \neq i} \frac{1}{m}\sum_{j \in C_k} d(i,j) \tag{2.2}$$

where m is the number of points in the cluster C_k.

We can define the silhouette score of one data point i as

$$s(i) = \frac{b(i)-a(i)}{\max(a(i),b(i))} \tag{2.3}$$

III. Simplified Evaluation Method for UL Models

In our study, we used a simplified evaluation method to evaluate the results of unsupervised learning models. This method is to calculate the mean or average values of testing data points and find the closest or similar mean values for the clustered data points to confirm and validate the unsupervised learning outcomes. To make things clear, we utilized a Diabetes dataset [15] as the input dataset to train and test some unsupervised learning algorithms, such as k-means, to confirm the effectiveness of our method.

During the training process, the top 500 observations of two features, Glucose and BMI, are used with k-means method to cluster them into two classes or clusters. Then the means for those two groups of observations are calculated based on their clusters

$$m_XC1 = \frac{1}{m}\sum_{i=1}^{m} X(idx == 1,:)$$
$$m_XC2 = \frac{1}{m}\sum_{i=1}^{m} X(idx == 2,:) \tag{3.1}$$

where m is the number of rows or observations (500) in the training data, and idx is the trained cluster number (1 or 2), and X is the training data (2 columns or features, Glucose and BMI).

Then the following 100 data in the Diabetes dataset with two columns, Glucose and BMI, work as the testing data XX to confirm the clustering results. The k-mean algorithm is used to cluster those testing data into two clusters (1 or 2), in which 1 means no diabetes (0 in the Outcome column in the Diabetes dataset) and 2 means diabetes (1 in the Outcome column in the Diabetes dataset). Then the mean values of each row or observation in the testing data with two columns are computed as:

$$m_XC = \frac{1}{n}\sum_{j=1}^{n}[XX(j,1) + XX(j,2)] \tag{3.2}$$

where n is the number of testing data (n = 100 in this study).

Finally, a comparison is performed to compare the mean values between the clustered training data, **m_XC1** and **m_XC2**, and the mean values of clustered testing data **m_XC**. If the difference between them is within a desired range, the clustering result is correct. Otherwise, they are incorrect. That desired error range needs to be experimentally determined based on real applications.

The operational procedure of using our method is shown below.

1) Training the k-means model by using the top 500 observations in Diabetes dataset with two columns, Glucose and BMI, to cluster them into 2 classes or clusters.
2) Calculate the mean values for two clustered observations, **m_XC1** and **m_XC2**.
3) Use another 100 observations in the Diabetes dataset as the testing records, and use k-means algorithm to cluster those 100 observations into 2 classes or clusters.
4) Calculate the mean or average values for 100 clustered testing data.
5) Compare the mean values of 100 clustered testing data with the mean values of the trained data, **m_XC1** and **m_XC2**, to confirm the matched cluster for all testing data.
6) Calculate the difference between the matched and unmatched clustered data in the testing data to get the clustering accuracy.

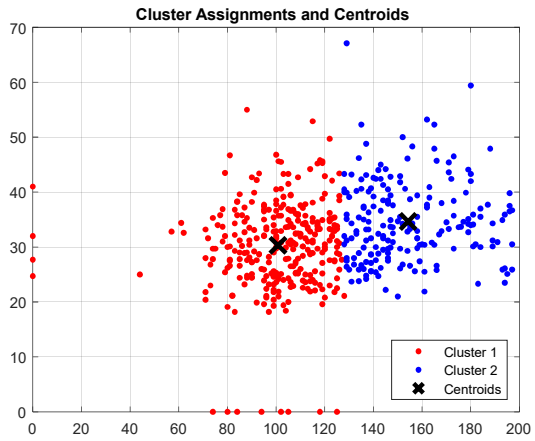


Fig. 1. The clustering result for top 500 observations on Diabetes dataset.

Figure 1 shows the clustering results for the top 500 training data. The horizontal is the Glucose data and the vertical indicates the BMI data.

IV A Comparison among Silhouette Score and Calinski-Harabasz Index with Our Simplified Method

A comparison study (Figure 2) is performed between the traditional Silhouette score method and our simplified method to compare the clustering accuracy for two methods. The last 100 records in our Diabetes dataset are used for this comparison purpose.

When using Silhouette score method, the clustering result is correct if the Silhouette score for the tested data is >= 0.5 and incorrect when that score is < 0.5. As for our simplified method, if the difference between the mean values of 100 clustered testing data and the mean values of the clustered training data is within a tolerance range, that clustering result can be considered as a correct one, otherwise it would be an incorrect one. This tolerance range needs to be determined experimentally by actual applications. In our study, this range is about 10.

First the Silhouette score method is used to validate the clustering results for those 100 data, and our simplified method is adopted to perform a similar validation for 100 data. The clustering correct rate or accuracy for the Silhouette score method is 77%, but the clustering correct rate for our simplified method is ranged between 77% and 97%.

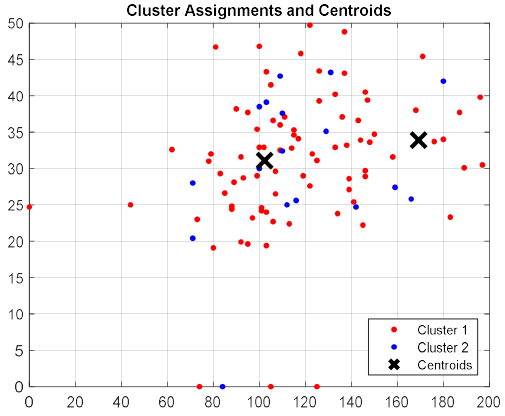
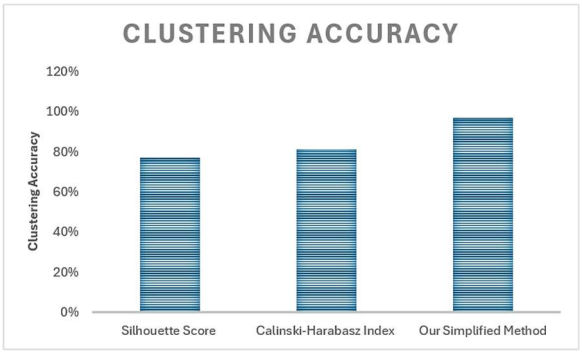


Fig. 2 Clustering result for the 100 observations on Diabetes dataset.

Table 1. Asmall collection of Silhouette Scores of top 10 testing data.

Number	Glucose	BMI	Cluster	Silhouette	Outcome	Result
1	117	25.2	1	0.7251	0	Correct
2	84	37.2	1	0.5612	0	Correct
3	96	39.0	1	0.5085	1	**Incorrect**
4	94	33.3	1	0.5189	0	Correct
5	96	37.3	1	0.8005	0	Correct
6	75	33.3	2	0.7063	0	**Incorrect**
7	180	36.5	2	0.8625	1	Correct
8	130	28.6	1	0.7397	0	Correct
9	84	30.4	2	0.7505	0	**Incorrect**
10	120	25.0	1	0.7396	0	Correct

Fig. 3 The histogram of the crusting accuracy for three methods.



Secondly, we also use the Calinski-Harabasz Index method to validate the clustering results for those 100 data. The clustering correct rate or accuracy for the Calinski-Harabasz Index method is about 81%. A histogram presentation of the crusting accuracy for three methods is shown in Figure 3.

Table 1 shows an example or a small collection of Silhouette Scores of top 10 testing data for two selected columns, Glucose and BMI. It can be found that the correct clustering rate is about 70%. In the Table 1, The column Cluster contained the clustering results for two features in the testing data, the column Outcome is the label for each row in the original Diabetes dataset. The purpose of using that column is to confirm and check the clustering results. A cluster 1 means no diabetes, which is equivalent to 0 in the Outcome column, and a cluster 2 indicates a diabetes case, which is equivalent to 1 in the Outcome column in the original Diabetes dataset. If a row in which the cluster is 1 and the corresponding Outcome is 0 with higher Silhouette score (≥ 0.5), which means that the clustering result is identical with that of Outcome value and the clustering result is correct. This is also true to the situation where a cluster is 2 and the corresponding Outcome is 1 with higher Silhouette score. Otherwise, if the cluster is 1 but the corresponding Outcome is also 1 with a higher Silhouette score, or if the cluster is 2 and the corresponding Outcome is 0 with higher Silhouette score, the clustering result would be wrong.

We built a MATLAB program to test and confirm our simplified evaluation algorithm, and the program running result shows below:

```
NumObservations: 500
    InspectedK: [1 2 3 4 5 6]
    CriterionValues: [NaN 0.7105 0.6008 0.5751 0.6045 0.5614]
    OptimalK: 2
ans =97
The correct clustering rate is 97
```

Due to some uncertainties existing in the K-Means clustering algorithm, the detected or evaluated correct clustering rate varies within a range, from 77% to 97%.

V CONCLUSION

A simplified evaluation algorithm used to effectively evaluate and validate unsupervised learning models is proposed in this study. Compared with some traditional validation metrics, such as Silhouette score and Calinski-Harabasz Index method, our method is better in the correct clustering rate, which is up to 97%, than the traditional validation metric's rate, which is 77% in the Silhouette score method.

REFERENCES

[1] https://www.proquest.com/openview/13e305db3d4d43c014a00e717dcbbf3f/1?pq-origsite=gscholar&cbl=18750.

[2] Julio-Omar Palacio-Niño and Fernando Berzal, "Evaluation Metrics for Unsupervised Learning Algorithms", arXiv eprint, May, 2019, https://arxiv.org/abs/1905.05667.

[3] Andreas Vlachos, "Evaluating unsupervised learning for natural language processing tasks", Proceedings of EMNLP 2011, Conference on Empirical Methods in Natural Language Processing, pages 35–42, Edinburgh, Scotland, UK, July 27–31, 2011.

[4] Garg, Vikas, and Adam T. Kalai. "Supervising unsupervised learning." Advances in Neural Information Processing Systems 31 (2018).

[5] Martin Q. Ma, Yue Zhao, Xiaorong Zhang, and Leman Akoglu. 2023. "The Need for Unsupervised Outlier Model Selection: A Review and Evaluation of Internal Evaluation Strategies". SIGKDD Explor. Newsl. 25, 1 (June 2023), 19–35. https://doi.org/10.1145/3606274.3606277.

[6] F. G. Portela, F. Almenares Mendoza and L. C. Benavides, "Evaluation of the performance of supervised and unsupervised Machine learning techniques for intrusion detection," 2019 IEEE International Conference on Applied Science and Advanced Technology (iCASAT), Queretaro, Mexico, 2019, pp. 1-8, doi: 10.1109/iCASAT48251.2019.9069538.

[7] G. Padmavathi, D. Shanmugapriya and A. Roshni, "Performance Analysis of Unsupervised Machine Learning Methods for Mobile Malware Detection," 2022 9th International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, 2022, pp. 201-206, doi: 10.23919/INDIACom54597.2022.9763180.

[8] T. T. Khoei, K. Al Shamaileh, V. K. Devabhaktuni and N. Kaabouch, "A Comparative Assessment of Unsupervised Deep Learning Models for Detecting GPS Spoofing Attacks on Unmanned Aerial Systems," 2024 Integrated Communications, Navigation and Surveillance Conference (ICNS), Herndon, VA, USA, 2024, pp. 1-10, doi: 10.1109/ICNS60906.2024.10550633.

[9] O. Toyin et al., "Comparative Analysis of Unsupervised Machine Learning Models on Early Detection of Breast Cancer," 2024 International Conference on Science, Engineering and Business for Driving Sustainable Development Goals (SEB4SDG), Omu-Aran, Nigeria, 2024, pp. 1-7, doi: 10.1109/SEB4SDG60871.2024.10630354.

[10] R. Barreto, L. Gomes and Z. Vale, "Smart Solution for Energy Communities: Integrating Demand Response and Unsupervised Learning Evaluation Metrics," 2024 IEEE 22nd Mediterranean Electrotechnical Conference (MELECON), Porto, Portugal, 2024, pp. 144-149, doi: 10.1109/MELECON56669.2024.10608742.

[11] A. U. Haq et al., "Detection of Breast Cancer Through Clinical Data Using Supervised and Unsupervised Feature Selection Techniques," in IEEE Access, vol. 9, pp. 22090-22105, 2021, doi: 10.1109/ACCESS.2021.3055806.

[12] Y. Zou, C. Liu, H. Shao and D. Zhong, "Unsupervised Palmprint Image Quality Assessment via Pseudo-Label Generation and Ranking Guidance," in IEEE Transactions on Instrumentation and Measurement, vol. 72, pp. 1-11, 2023, Art no. 5018811, doi: 10.1109/TIM.2023.3288259.

[13] https://fiveable.me/modern-statistical-prediction-and-machine-learning/unit-14/evaluation-metrics-unsupervised-learning-clustering/study-guide/UuzjLvrGcbnj7uhw.

[14] https://medium.com/@MrBam44/how-to-evaluate-the-performance-of-clustering-algorithms-3ba29cad8c03.

[15] https://www.kaggle.com/datasets/mathchi/diabetes-data-set.

Copyright © 2026 by the authors. This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited (CC BY 4.0).