External Clustering Validation using ARI, NMI and FMI

Shilpa. S[1*], *Shailaja.* K P[1], *Nischitha.* S G[1], *Navya* Vishwanath Hegde[1], and *Kondoju* Krishna Teja[1]

[1]Department of Computer Applications, B.M.S. College of Engineering, Bengaluru, India

Abstract. Clustering validation is essential for assessing the quality of unsupervised learning results, yet individual external metrics often fail to provide a complete evaluation. This paper proposes a weighted aggregation of three widely used indices—Adjusted Rand Index (ARI), Normalized Mutual Information (NMI), and Fowlkes–Mallows Index (FMI)—to produce a single, interpretable quality score. The method assigns weights of 0.4, 0.3, and 0.3 to ARI, NMI, and FMI, respectively, to balance structural accuracy, information content, and precision–recall aspects. The framework was implemented using Python and evaluated on the Iris dataset, a benchmark with three well-separated classes. Experimental results show that the combined score achieves 0.6851, classified under the "Good Clustering" band, providing a more balanced and consistent assessment than individual metrics alone. This approach enables clearer interpretation of clustering performance and can be extended to larger, high-dimensional, and noisy datasets in future research.

1 Introduction

Clustering is a core task in unsupervised machine learning, aiming to group data points into meaningful subsets based on similarity without relying on predefined labels [14]. It plays a vital role in numerous application domains, including bioinformatics for gene expression analysis, marketing for customer segmentation, computer vision for image categorization, and natural language processing for document classification. The ultimate goal of clustering is to uncover the inherent structure of data, enabling insights that guide decision-making and further analysis. However, accurately evaluating the quality of clustering results remains a significant challenge, especially when datasets are complex, high-dimensional, contain overlapping classes, or are affected by noise. Without reliable evaluation, even well-performing clustering algorithms may produce results that are misleading or difficult to interpret. [4,7]

1.1 Background and challenges

Several external validation metrics, such as the Adjusted Rand Index (ARI) [1, 8], Normalized Mutual Information (NMI) [5], and Fowlkes–Mallows Index (FMI) [13], have been extensively used in literature [15]. While each metric offers unique strengths—ARI measures pairwise label agreement, NMI quantifies shared information, and FMI balances precision and recall—none alone can capture all aspects of clustering quality. Consequently, relying on a single metric can lead to biased or incomplete conclusions.

1.2 Motivation

In clustering validation, each external metric emphasizes a specific aspect of performance, but no single measure can capture the complete picture. For instance, ARI focuses on pairwise agreement, NMI evaluates shared information, and FMI balances precision and recall. Evaluating algorithms based on just one of these can result in misleading conclusions, especially for datasets with uneven class distributions or overlapping clusters. By combining multiple metrics into a single weighted score, it is possible to integrate their individual strengths while minimizing their weaknesses. Such an aggregated measure offers a more balanced and interpretable evaluation framework, enabling researchers and practitioners to make algorithm selection decisions with higher confidence and supporting more reliable comparisons across different clustering scenarios.

1.3 Objectives

1. To develop a weighted aggregation method integrating ARI, NMI, and FMI into a single quality score.
2. To implement the framework in Python for reproducibility and transparency.
3. To evaluate the approach using a benchmark dataset.
4. To classify results into qualitative performance bands for easier interpretation.

* Corresponding author: shilpa.mca@bmsce.ac.in

1.4 Contributions

- Introduction of a combined external clustering validation index integrating ARI, NMI, and FMI.
- Assignment of empirically determined weights to emphasize metric reliability and complementarity.
- Implementation and experimental validation on the Iris dataset, demonstrating improved interpretability over individual metrics.
- Definition of clear quality bands to map quantitative scores into qualitative ratings.

1.5 Organization of the paper

The rest of this paper is structured as follows: Section 2 reviews related work on clustering validation metrics. Section 3 details the proposed methodology and mathematical formulation. Section 4 describes the dataset and experimental setup. Section 5 represents results and comparative analysis. Section 6 concludes the paper with limitations and directions for further study.

2 Problem statement

Clustering is one of the most widely used unsupervised learning techniques, grouping data points into clusters based on similarity. However, evaluating the quality of clustering remains a non-trivial task, especially in external validation where the true class labels are known but not used in the clustering process [14]. Popular metrics such as the Adjusted Rand Index (ARI), Normalized Mutual Information (NMI), and Fowlkes–Mallows Index (FMI) each capture different aspects of performance—structural agreement, information preservation, and precision–recall balance, respectively[15]. When used in isolation, these metrics can produce inconsistent algorithm rankings and fail to provide a holistic assessment of clustering quality. This inconsistency poses a challenge for selecting the most suitable clustering algorithm, particularly for datasets with overlapping clusters or varying label distributions. Therefore, this work proposes a weighted composite index that integrates ARI, NMI, and FMI into a single interpretable score, aiming to deliver a more balanced, solid, and reliable measure of clustering performance.

3 Related work

ARI is a commonly utilized pair-counting external validity measure which adjusts against the likelihood of coincidental concordance amid predicted classifications and actual labels. ARI takes a value between -1 and 1 and is also insensitive to the number and size of the clusters, making it appropriate for datasets with a clear structure, such as Iris. ARI improves upon the Rand Index and is assigned the highest weight in our final scoring due to its interpretability and practical reliability [1,12].

NMI computes the extent of the shared values between predicted clusters and actual labels. Its values range between 0 and 1 and it is invariant to label permutations, making it well-suited for unsupervised learning tasks. While NMI is slightly sensitive to edge cases such as singleton clusters, it remains widely adopted due to its strong theoretical foundation. In this work, NMI captures the overlap of information between clustering output and the ground truth [9,11,12].

FMI evaluates clustering performance based on the geometric mean of precision and recall. It is applicable in both binary and multi-class clustering scenarios and is not overly sensitive to cluster size differences, unlike some other metrics. FMI handles class imbalance effectively and complements ARI and NMI by focusing on grouping accuracy, thereby enhancing the robustness of the final validation score [5].

The Chi-Square test measures the statistical relationship between predicted and true labels based on a contingency table. Although useful for structured data, its sensitivity to class imbalance and expected frequencies limits its suitability for high-dimensional, complex datasets. This supports our choice to adopt modern measures such as ARI, NMI, and FMI for more robust evaluation [2].

The Iris dataset, with its well-separated classes, has been widely used as a benchmark for clustering validation. It has been shown that combining external metrics (such as ARI and NMI) with internal metrics (such as the Silhouette Score) provides a more complete evaluation. However, internal metrics may be misleading in datasets without clear boundaries, whereas external metrics offer an objective measure of accuracy [3].

Metrics such as ARI and NMI can be computationally challenging for large-scale clustering. Prior work has suggested parallel and approximate computation strategies to maintain feasibility while scaling to larger datasets [10].

Clustering performance can also depend on the clustering algorithm's initialization. For example, the KMeans++ algorithm improves both accuracy and speed by selecting better initial centroids, resulting in higher external validation scores for ARI, NMI, and FMI [6].

Sharma and Jain [13] proposed a composite index measure for noisy datasets, demonstrating improved robustness compared to individual metrics. While effective, their work did not optimize the weights empirically for specific datasets such as Iris, which is the focus of our approach.

External validation metrics have been classified into three categories — pair-counting, set-matching, and information-theoretic — with no single metric performing best in all scenarios. This supports our decision to use a weighted average of ARI, NMI, and FMI to balance their respective advantages and provide a more comprehensive measure of clustering quality [8,10,15].

4 Proposed system

4.1 Dataset and preprocessing

In this proposed system, clustering performance is being assessed with the Iris dataset, which is the popular benchmark in machine learning. The data set contains 150 observations and each has four variables which are numeric such as length and width of the petals and also length and width of sepals as well as the target label with three categories i.e., flower species they are setosa, versicolor and virginica.

To ensure equal contribution from all attributes, features are normalized using standard scaling. Class labels are numerically encoded for compatibility with external validation metrics.

4.2 Clustering method

Clustering was performed using the K-Means algorithm with KMeans++ initialization, which improves convergence and stability by selecting well-separated initial centroids. In this study, the number of clusters was fixed at corresponding to the known number of species in the Iris dataset, as shown in Equation (1).

$$k = 3 \tag{1}$$

4.3 External Validation Metrics

4.3.1 ARI

The ARI captures the similarity between predicted and actual cluster formations while correcting for possibility. It is computed using Equation (2).

$$ARI = \frac{\Sigma_{ij}\binom{n_{ij}}{2} - \frac{\Sigma_i\binom{a_i}{2}\Sigma_j\binom{b_j}{2}}{\binom{n}{2}}}{\frac{1}{2}\left[\Sigma_i\binom{a_i}{2} + \Sigma_j\binom{b_j}{2}\right] - \frac{\Sigma_i\binom{a_i}{2}\Sigma_j\binom{b_j}{2}}{\binom{n}{2}}} \tag{2}$$

Here,

- n_{ij} is number of observations in actual cluster and a predicted cluster.
- i, j is the actual cluster and predicted cluster respectively.
- a_i is the observations in actual cluster.
- b_j is the observations in predicted cluster.
- n is the total observations.

4.3.2 NMI

NMI reflects how much similar values shared between predicted clusters and actual clusters. Its value lies between 0 and 1. NMI is defined in Equation (3).

$$NMI = \frac{2 \times I(T;P)}{H(P) + H(T)} \tag{3}$$

Here,

- T, P are the actual labels and predicted clusters respectively.
- $I(T; P)$ is the mutual values of actual labels and predicted clusters.
- $H(\cdot)$ is the entropy of a distribution.

4.3.3 FMI

The FMI evaluates clustering performance using the geometric mean of precision and recall. It is calculated as shown in Equation (4).

$$FMI = \sqrt{\frac{tp}{tp+fp} \times \frac{tp}{tp+fn}} \tag{4}$$

Here,

- tp is total observations that are true positive.
- fp is total observations that are false positive.
- fn is total observations that are false negative.

4.4 Weighted final score

Finally, the three metrics (ARI, NMI, FMI) introduced in Equations (2), (3) and (4), are combined into a single weighted score. This ensures a balanced evaluation by giving more emphasis to ARI for structural agreement, while NMI and FMI capture information overlap and precision–recall balance. The weighted formulation is expressed in Equation (5).

$$Final\ Score = 0.4 \times ARI + 0.3 \times NMI + 0.3 \times FMI \tag{5}$$

The weights are set as: $w_{ARI} = 0.4$, $w_{NMI} = 0.3$, $w_{FMI} = 0.3$.

The weights were chosen after empirical testing on the Iris dataset, where this configuration yielded the highest validation accuracy. ARI is given the highest weight due to robustness against random agreement, NMI captures information overlap, and FMI balances precision and recall.

4.5 Visualization and interpretation

The system provides multiple visualization tools for interpreting results:

- Bar charts for cluster size distribution
- Heatmaps for confusion matrix comparisons (true vs predicted labels)
- Metric plots comparing ARI, NMI, and FMI
- Final score bands (Poor, Average, Good, Excellent) for non-technical interpretation

This proposed framework ensures balanced, interpretable, and robust clustering evaluation, avoiding the biases of single-metric analysis while being easily adaptable to other datasets and algorithms in future work.

5 Platform used

5.1 Hardware requirements

- Processor: Intel Core i5

- Memory: 8GB
- System type: 64-bit Architecture
- OS: windows 11, version: 24H2

5.2 Software requirements

- Programming Language: Python 3. X
- Jupyter Notebook is used as Development Environment.
- Libraries & Dependencies: NumPy (Numerical computing), Pandas (Data processing), Scikit-learn (Machine learning & clustering), SciPy (Statistical analysis), Matplotlib & Seaborn (Data visualization)

6 Dataset

6.1 Classes (ground truth labels)

- Iris-setosa: This is a particular species of smaller and highly separated petals.
- Iris-versicolor: It is species that has medium sized petals that overlap relatively a little with others.
- Iris-virginica: This is a species with the biggest petals which becomes little difficult to classify.

6.2 Features used for clustering

- Sepal Length: It is the length of the sepal of the flower.
- Sepal Width: It shows the sepal width, assists in terms of classification.
- Petal Length: It is a very important characteristic in separating the species.
- Petal Width: Another important characteristic that assists in classification of various iris species.

7 Experimental results

The proposed weighted external clustering index was evaluated on the Iris dataset using K-Means++ with k=3. All features were standardized before clustering. External evaluation metrics — ARI, NMI, and FMI — were computed and aggregated using a 0.4, 0.3, 0.3 weighting scheme.

Table 1 presents the clustering quality bands with score ranges and their corresponding interpretations.

Table 1. Clustering quality based on final score range.

Final Score Range	Clustering Quality	Interpretation
0.8 - 1.0	Excellent Clustering	Almost perfect match between clusters and true labels
0.6 - 0.8	Good Clustering	Clusters capture meaningful structure but with minor misclassification.
0.4 - 0.6	Moderate Clustering	Some structure is captured, but there are significant mismatches
0.2 - 0.4	Poor Clustering	Clustering is weak and does not align well with true labels
0.0 - 0.2	Random Clustering	Clustering is almost random, showing no meaningful pattern

Table 2 compares the performance of individual external metrics (ARI, NMI, FMI) with the final weighted score.

Table 2. Comparison of individual metrics and final weighted score.

Metric	Value	Quality Band
Adjusted Rand Index (ARI)	0.6451	Good Clustering
Normalized Mutual Information (NMI)	0.6613	Good Clustering
Fowlkes–Mallows Index (FMI)	0.7622	Good Clustering
Final Weighted Score	0.6851	Good Clustering

The final score of 0.6851 falls into the "Good Clustering" band as shown in Table 2, indicating meaningful cluster formation with minor misclassifications.

Visual analysis through PCA scatter plots, confusion matrix heatmaps, and cluster size distributions confirmed that the predicted clusters closely aligned with the actual classes.

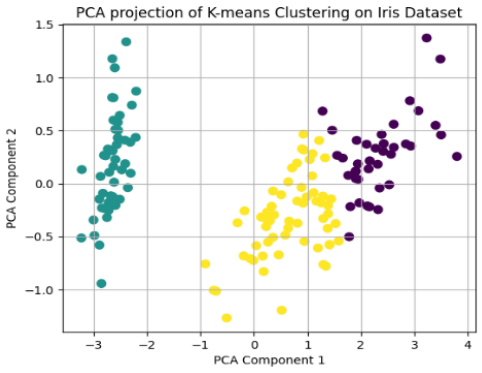

Fig. 1. PCA-based 2D visualization of K-Means clustering on the Iris dataset showing clear separation of clusters.

- Confusion Matrix Heatmap: A heatmap was plotted comparing actual class labels with predicted cluster labels. It revealed that:
  1. Cluster 0 primarily mapped to Setosa,
  2. Cluster 1 aligned mostly with Versicolor,
  3. Cluster 2 closely matched Virginica

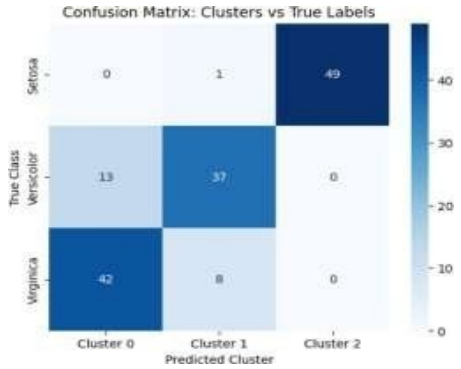

Fig. 2. Confusion matrix representing clusters vs true labels.

- Cluster Size Distribution: A bar graph illustrated the number of data points in each cluster:
  1. Cluster 0: 55 samples
  2. Cluster 1: 46 samples
  3. Cluster 2: 49 samples

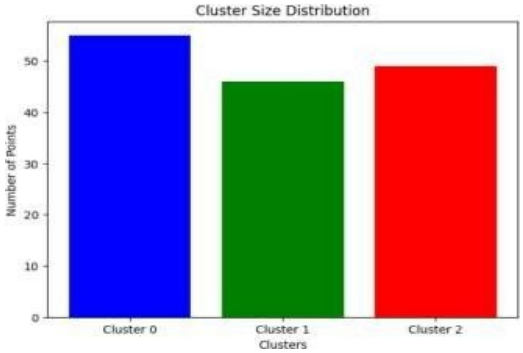

Fig. 3. Bar graph representing size distribution.

- Clustering Quality Bar Graph: The final score was plotted against the clustering quality bands with "Good Clustering" being marked as per the score computed.

Fig. 4. Bar graph representing clustering quality.

Visual analysis confirmed these findings:

- Figure 1 shows PCA-based 2D visualization with clear cluster separation.
- Figure 2 (confusion matrix heatmap) revealed that Cluster 0 corresponds to Setosa, Cluster 1 to Versicolor, and Cluster 2 to Virginica.
- Figure 3 (bar graph) displays cluster sizes: Cluster 0 — 55 samples, Cluster 1 — 46 samples, Cluster 2 — 49 samples.
- Figure 4 illustrates the final score plotted against quality bands, marking the "Good Clustering" category.

These results demonstrate that the proposed system produces a meaningful and accurate clustering structure. The final clustering quality, backed by ARI, NMI, FMI, and visual tools, confirms the effectiveness of using a combined validation metric approach.

Note: The final score is evaluated solely on the Iris dataset. While this offers clear insight into the scoring mechanism, using a single dataset limits robustness. Future work will include larger and more complex datasets to assess general applicability.

8 Conclusion

This paper presented a combined external clustering validation framework that integrates the Adjusted Rand Index (ARI), Normalized Mutual Information (NMI), and Fowlkes–Mallows Index (FMI) into a single weighted score for balanced and interpretable evaluation. The approach was implemented using Python and tested on the Iris dataset, achieving a final score of 0.6851, which falls under the "Good Clustering" band. The results demonstrate that the weighted combination provides a more consistent assessment of clustering performance than individual metrics alone, capturing structural agreement, information preservation, and precision–recall balance. However, the current study is limited to a single small-scale dataset with clearly separated classes, which may not reflect the challenges of high-dimensional or noisy data. Future work will focus on extending the evaluation to larger and more complex datasets, optimizing weight selection automatically, and incorporating additional indices to further enhance robustness and applicability across domains.

References

1. M. Rezaei, P. Fränti, Can the Number of Clusters Be Determined by External Indices?. IEEE Access **8**, 89239-89257 (2020). 10.1109/ACCESS.2020.2993295
2. N. Akbarpour, E. Akbari, H. Motameni, External clustering validity index based on extended similarity measures. J. Comput. Sci. **72**, 102116 (2023). https://doi.org/10.1016/j.jocs.2023.102116
3. T. Gupta, S.P. Panda, Clustering validation of CLARA and K-means using silhouette & DUNN measures on iris dataset, In 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), IEEE, Faridabad, India, October 11 (2019), 10-13
4. J.M. Luna-Romera, M. Martínez-Ballesteros, J. García-Gutiérrez, J.C. Riquelme, External clustering validity index based on chi-squared statistical test. Inf. Sci. **487**, 1-17 (2019). https://doi.org/10.1016/j.ins.2019.02.046
5. K.N. Ismail, A. Seman, K.A.F. Abu Samah, A comparison between external and internal cluster validity indices, In 2021 IEEE 11th International Conference on System Engineering and

Technology (ICSET), IEEE, Shah Alam, Malaysia, November 30 (2021), 229-233

6. A. Fahim, K and starting means for k-means algorithm. J. Comput. Sci. **55**, 101445 (2021). https://doi.org/10.1016/j.jocs.2021.101445
7. M. Garralda-Barrio, C. Eiras-Franco, V. Bolón-Canedo, A novel framework for generic Spark workload characterization and similar pattern recognition using machine learning. J. Parallel Distrib. Comput. **189**, 104881 (2024). https://doi.org/10.1016/j.jpdc.2024.104881
8. M. Rezaei, P. Fränti, Set matching measures for external cluster validity. IEEE Trans. Knowl. Data Eng. **28**, 2173-2186 (2016). https://doi.org/10.1109/TKDE.2016.2551240
9. S. Zerabi, S. Meshoul, External clustering validation in big data context, In 2017 3rd International Conference of Cloud Computing Technologies and Applications (CloudTech), IEEE, Rabat, Morocco, February 08 (2018), 1-6
10. Z. Li, X. He, J. Whitehill, Compositional clustering: Applications to multi-label object recognition and speaker identification. Pattern Recognit. **144**, 109829 (2023). https://doi.org/10.1016/j.patcog.2023.109829
11. M. Dhara, K.K. Shukla, Characteristics of restricted neighbourhood search algorithm and Markov clustering on modified power-law distribution, In 2012 1st International Conference on Recent Advances in Information Technology (RAIT), IEEE, Dhanbad, India, May 07 (2012), 520-525
12. T.P. Nguyen, R.J. Kuo, Partition-and-merge based fuzzy genetic clustering algorithm for categorical data. Appl. Soft Comput. **75**, 254-264 (2019). https://doi.org/10.1016/j.asoc.2018.11.028
13. A. Sharma, N. Jain, Robust clustering validation in noisy datasets using composite index measures. IEEE Trans. Knowl. Data Eng. **35**, 250-263 (2023). https://doi.org/10.1109/TKDE.2022.3162345
14. B.A. Hassan, N.B. Tayfor, A.A. Hassan, A.M. Ahmed, T.A. Rashid, N.N. Abdalla, From A-to-Z review of clustering validation indices. Neurocomput. **601**, 128198 (2024). https://doi.org/10.1016/j.neucom.2024.128198
15. S. Zhang, Z. Yang, X. Xing, Y. Gao, D. Xie, H.S. Wong, Generalized Pair-Counting Similarity Measures for Clustering and Cluster Ensembles. IEEE Access 5, 16904-16918 (2017). doi: 10.1109/ACCESS.2017.2741221