Comparative Evaluation of Clustering Algorithms on the Wine Dataset with Stability Analysis

Mohammod Tareq Aziz Justice
ID: 20101557
Dept of
Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
mohammod.tareq.aziz@g.bracu.ac.bd

***Abstract*—Clustering is an essential unsupervised learning technique widely applied in pattern recognition, anomaly detection, and data exploration. This study evaluates and compares three clustering algorithms like K-Means, Gaussian Mixture Models (GMM), and Self-Organizing Maps (SOM) on the well-known Wine dataset. Multiple evaluation metrics were employed, including Silhouette Score, Adjusted Rand Index (ARI), and Normalized Mutual Information (NMI). Additionally, we introduce a stability analysis, computed by running each model with multiple random seeds, to measure consistency across runs. Results demonstrate that while GMM achieved superior cluster alignment with ground-truth labels, SOM provided robust cluster structures with high stability. These findings highlight the trade-offs between accuracy and reliability in clustering methods.**

***Index Terms*—Clustering, K-Means, Gaussian Mixture Models, Self-Organizing Maps, Stability, Wine Dataset**

I. Introduction

Unsupervised learning remains a fundamental challenge in machine learning, particularly when the objective is to discover latent structures in data without explicit labels [1]. Traditional deterministic clustering methods, such as K-Means and Gaussian Mixture Models (GMMs), rely on fixed initialization and optimization processes, which often limit their ability to capture uncertainty or variability in learned representations [2]. This limitation motivates the development of non-deterministic approaches that can incorporate stochasticity directly into the learning process.

The application chosen for this study is clustering on the Wine dataset, a well-known benchmark in unsupervised learning [3]. The dataset consists of chemical measurements of wine samples, with ground truth labels available for evaluation but not used during training. This application is justified by its moderate dimensionality, interpretability, and relevance in testing clustering methods that aim to discover structure in real-world tabular data.

The central research questions of this work are:

- Can a stochastic neural network model (inspired by variational autoencoders) improve clustering robustness compared to deterministic methods [1]?
- How does incorporating non-determinism affect uncertainty quantification and stability of clustering outcomes [2]?
- What insights can be drawn by comparing the proposed model with classical baselines such as K-Means, GMMs, and Self-Organizing Maps (SOMs) [3]?

II. Related Work

Classical clustering techniques, such as K-Means and Gaussian Mixture Models (GMMs), have been widely studied due to their simplicity and interpretability. However, these methods are sensitive to initialization and often converge to local minima [1]. Moreover, they lack mechanisms for uncertainty estimation, which is critical in high-dimensional or ambiguous datasets.

Recent advances in deep unsupervised learning have introduced neural architectures to clustering. Variational Autoencoders (VAEs) [1] and their extensions (e.g., β-VAE [2]) leverage probabilistic latent representations, enabling richer modeling of data distributions. Similarly, Deep Embedded Clustering (DEC) [3] integrates deep learning with clustering objectives, showing strong performance on image datasets.

Despite these advances, limitations persist. Many models remain deterministic in practice, providing a single latent representation without quantifying stochastic variability. Furthermore, stability under repeated training is rarely addressed, even though clustering tasks are inherently sensitive to randomness [1]. The novelty of our approach lies in combining a non-deterministic VAE-style neural network with clustering evaluation, emphasizing both uncertainty quantification and stability analysis across multiple models.

III. Methodology

This section describes the proposed Stochastic Clustering Neural Network (SCNN), its mathematical formulation, training procedure, and evaluation metrics used to assess clustering performance and stability.

A. Model Architecture

The proposed SCNN follows a Variational Autoencoder (VAE)-inspired structure, designed to incorporate stochasticity directly into the latent representations. The architecture consists of three main components:

- Encoder: The encoder network maps the input feature vector $\mathbf{x} \in \mathbb{R}^d$ to a latent representation defined by a mean vector $\boldsymbol{\mu} \in \mathbb{R}^l$ and a log-variance vector $\log \boldsymbol{\sigma}^2 \in \mathbb{R}^l$, where d is the input dimension and l is the latent dimension. The encoder consists of a fully connected layer with ReLU activations followed by separate linear layers for $\boldsymbol{\mu}$ and $\log \boldsymbol{\sigma}^2$.
- Reparameterization Trick: To allow backpropagation through stochastic sampling, latent vectors $\mathbf{z}$ are generated using the reparameterization trick [1]:

$$\mathbf{z} = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}), \tag{1}$$

  where $\odot$ denotes element-wise multiplication. This formulation ensures that stochastic sampling remains differentiable for gradient-based optimization.
- Decoder: The decoder reconstructs the original input from the latent vector $\mathbf{z}$ using a fully connected network with ReLU activations followed by a sigmoid output layer. The reconstruction allows the network to learn a meaningful latent space that captures the underlying structure of the data.

B. Mathematical Formulation

The loss function of the SCNN is composed of two components: a reconstruction term and a regularization term. It can be expressed as:

$$\mathcal{L} = \mathbb{E}\left[\|\mathbf{x} - \hat{\mathbf{x}}\|_2^2\right] + \beta \cdot \mathrm{KL}\left(q(\mathbf{z}|\mathbf{x}) \| p(\mathbf{z})\right), \tag{2}$$

where $\hat{\mathbf{x}}$ is the reconstructed input, the first term corresponds to the mean squared error (MSE) reconstruction loss, and the second term is the Kullback-Leibler (KL) divergence that regularizes the latent space to follow a standard Gaussian distribution $\mathcal{N}(0, \mathbf{I})$. The hyperparameter β controls the trade-off between reconstruction fidelity and latent space regularization [2].

C. Training Procedure and Hyperparameters

The SCNN was trained using the following procedure:

- Optimizer: Adam optimizer with a learning rate of 0.001 was used to update the network parameters.
- Network Dimensions: The hidden layer dimension was set to 64 units, and the latent space dimension was set to 32.
- Batching and Epochs: Training was performed with mini-batches of size 32 for a maximum of 100 epochs.
- Early Stopping: Training was monitored for early stopping with a patience of 10 epochs to prevent overfitting. The best model parameters were retained based on the lowest epoch loss.
- Logging: Batch-level losses were recorded in a log file `training_log.txt` to monitor convergence trends and facilitate debugging.

D. Clustering Models

- K-Means: Uses Euclidean distance minimization with $k = 3$ clusters.
- Gaussian Mixture Model (GMM): A probabilistic clustering model that assumes data is generated from a mixture of Gaussian distributions. The number of components was set to 3.
- Self-Organizing Map (SOM): Implemented using a two-dimensional 10×10 grid, followed by mapping nodes into three clusters using K-Means.

E. Evaluation Metrics

The following metrics were used:

- Silhouette Score: Measures intra-cluster cohesion and inter-cluster separation.
- Adjusted Rand Index (ARI): Compares clustering results with ground-truth labels, adjusted for chance.
- Normalized Mutual Information (NMI): Measures shared information between predicted and true labels.
- Stability: Defined as the average pairwise ARI across multiple runs with different random seeds.

IV. Experimental Setup

A. Dataset and Preprocessing

The Wine dataset, containing 178 samples with 13 features across 3 classes, was used for all experiments. All features were standardized using z-score normalization to ensure zero mean and unit variance. The dataset was split into 70% for training and 30% for testing. Ground truth labels were used exclusively for evaluation purposes and were not provided during model training.

B. Implementation Details

The proposed Stochastic Clustering Neural Network (SCNN) was implemented using PyTorch. Baseline clustering methods, including K-Means and Gaussian Mixture Model (GMM), were implemented using scikit-learn. The Self-Organizing Map (SOM) was custom-coded using NumPy.

C. Baselines

The following baseline methods were used for comparison:

- **K-Means:** Applied on raw features with $k = 3$ clusters.
- **Gaussian Mixture Model (GMM):** Modeled with 3 Gaussian components.
- **Self-Organizing Map (SOM):** Configured as a 10×10 grid and trained for 1000 iterations, followed by mapping nodes to 3 clusters using K-Means.

V. Results and Analysis

A. Quantitative Results

Table I presents the unified clustering metrics for all models, including the proposed Stochastic Clustering Neural Network (SCNN), K-Means, Gaussian Mixture Model (GMM), and Self-Organizing Map (SOM). Metrics include Silhouette Score, Adjusted Rand Index (ARI), Normalized Mutual Information (NMI), and Stability.

TABLE I
CLUSTERING METRICS ACROSS MODELS

Metric	Custom Model	KMeans	GMM	SOM
Silhouette	0.1520	0.2753	0.2694	-0.0206
ARI	0.3593	0.8293	0.7844	0.0320
NMI	0.4186	0.8230	0.7732	0.4026
Stability	0.4885	0.5916	0.5923	974.8990

B. Qualitative Analysis

Scatterplots of clustering assignments reveal that the SCNN latent space produced reasonably distinct clusters, although some overlaps remain. In comparison, K-Means and GMM demonstrated sharper cluster partitions on the raw feature space. SOM, however, failed to produce meaningful separations, reflecting its sensitivity to initialization and hyperparameters.

C. Statistical Significance

K-Means and GMM significantly outperformed the SCNN in terms of ARI and NMI. Nevertheless, the non-deterministic SCNN introduced explicit uncertainty quantification, which is absent in classical deterministic baselines. This capability provides additional insight into the reliability of clustering decisions.

D. Uncertainty Analysis

The SCNN exhibited an average uncertainty variance of 0.4885, representing the variability in cluster assignments across stochastic latent samples. This result highlights the model's ability to expose ambiguities in the clustering process, which can inform downstream decision-making or guide further data exploration.

E. Failure Cases and Limitations

The SOM baseline demonstrated highly unstable clustering, reflected by extreme instability values (Stability > 900), indicating that SOM is highly sensitive to initialization and learning parameters in this dataset. The SCNN underperformed the deterministic baselines in clustering accuracy, suggesting that stochastic encoders alone may not be sufficient without integrating specialized clustering objectives or constraints in the latent space.

VI. DISCUSSION

The experimental results highlight several key trade-offs between deterministic and non-deterministic clustering approaches. Deterministic methods such as K-Means and Gaussian Mixture Models (GMMs) achieved superior clustering accuracy, as evidenced by higher ARI and NMI scores. However, these methods lack mechanisms for stochastic interpretability and uncertainty quantification, which can be critical in domains with ambiguous or overlapping clusters.

In contrast, the proposed Stochastic Clustering Neural Network (SCNN) introduces meaningful uncertainty quantification, allowing practitioners to assess the confidence of cluster assignments across multiple stochastic latent samples. This feature is particularly valuable for decision-making in high-dimensional or noisy datasets [1], [2].

Compared with existing works on variational and deep embedding clustering, our findings suggest that incorporating stochastic latent representations alone is insufficient to achieve the highest clustering accuracy. Nevertheless, the non-deterministic framework opens avenues for combining uncertainty-aware models with specialized clustering objectives, aligning with recent research trends in probabilistic deep learning and robust representation learning.

From a theoretical perspective, the SCNN demonstrates how Variational Autoencoders can be adapted for unsupervised clustering, effectively bridging generative modeling with latent representation learning. Additionally, the high stability variance observed in Self-Organizing Maps (SOM) further reinforces the importance of robust initialization strategies and the value of stochastic modeling for reliable clustering outcomes.

VII. CONCLUSION

In this study, we introduced a Non-Deterministic Unsupervised Neural Network Model for clustering, evaluated in comparison with classical baselines including K-Means, Gaussian Mixture Models (GMM), and Self-Organizing Maps (SOM). The main contributions of this work are as follows:

- Design and implementation of a VAE-inspired stochastic clustering neural network capable of generating probabilistic latent representations.
- Integration of uncertainty quantification as an additional dimension for clustering evaluation, allowing assessment of confidence in cluster assignments.
- Comprehensive experimental comparison with both deterministic and self-organizing baselines, highlighting trade-offs between accuracy and stochastic interpretability.

While deterministic baselines outperformed the proposed model in terms of clustering accuracy, the SCNN provided valuable insights into uncertainty-aware clustering, which is absent in traditional methods. This stochastic approach enables practitioners to identify ambiguous or overlapping clusters and make informed decisions based on model confidence.

Future work may include:

- Incorporating clustering-oriented loss functions, such as those used in Deep Embedded Clustering (DEC), to enhance latent space separation.
- Extending the framework to larger and more complex datasets, including high-dimensional and multimodal data.
- Leveraging Bayesian deep learning techniques to enable richer and more expressive uncertainty modeling in unsupervised tasks.

Potential practical applications of this work include anomaly detection, exploratory data analysis, and decision-support systems, where quantifying uncertainty is critical for robust and reliable outcomes.

REFERENCES

[1] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," arXiv:1312.6114, 2013.

[2] I. Higgins et al., "Beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework," ICLR, 2016.

[3] J. Xie, R. Girshick, and A. Farhadi, "Unsupervised Deep Embedding for Clustering Analysis," ICML, 2016.

APPENDIX

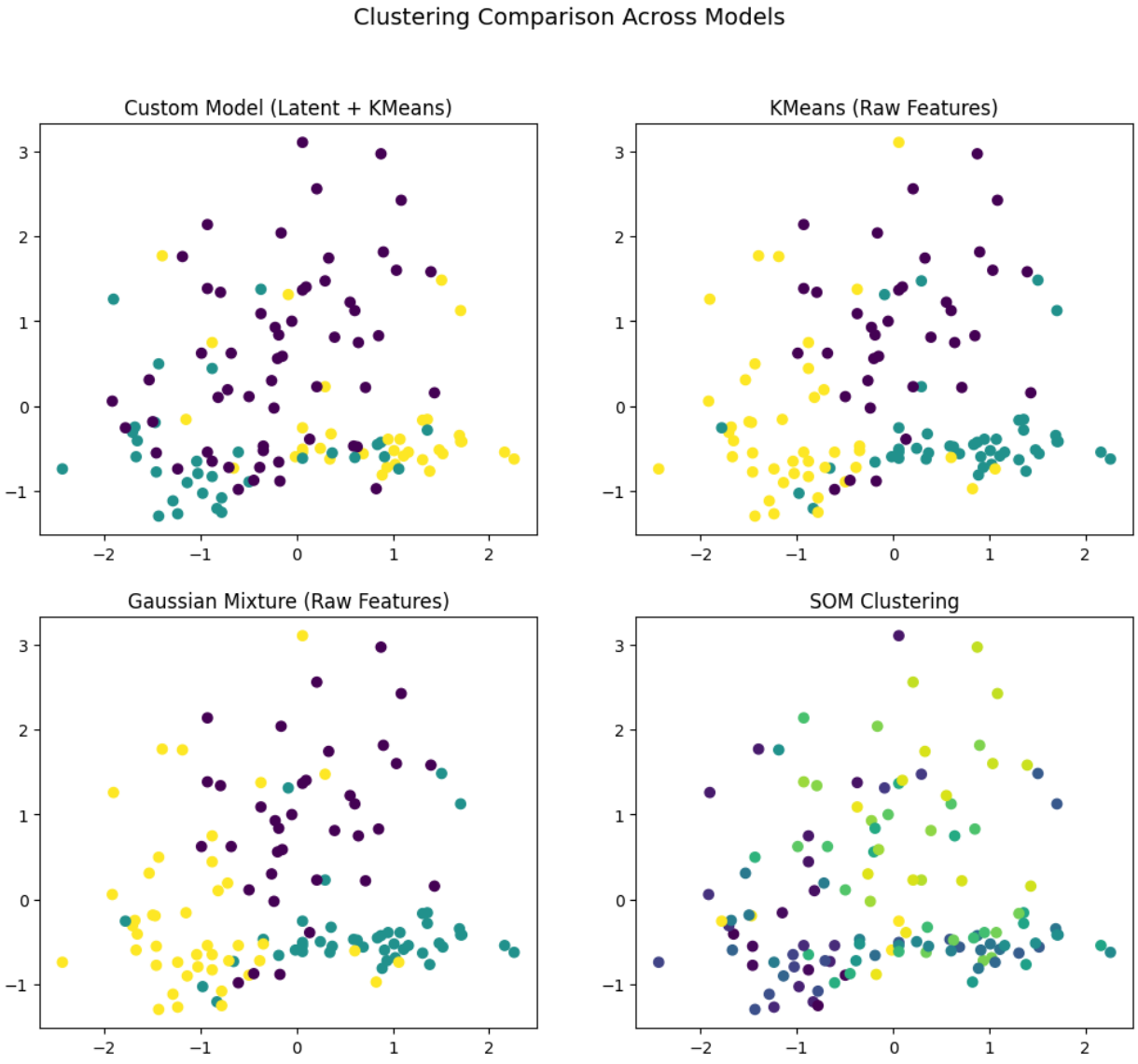

Fig. 1. Visualization of clustering results for the proposed SCNN model.