Counterfactual Distances for Clustering Quality and Separation-Aware Clustering

Georgios Vardakas, Antonia Karra, Evaggelia Pitoura, Aristidis Likas
Department of Computer Science and Engineering, University of Ioannina, Greece

Abstract

• We introduce a counterfactual-based formulation of cluster separation, where separation is measured as the distance of points to decision boundaries of competing clusters.
• We propose **Counterfactual-based Quality Score** (CFQ), an internal clustering quality index for centroid-based clustering that combines counterfactual separation with intra-cluster variance.
• Empirical evaluation shows that CFQ provides a robust alternative to Silhouette and VRC, capturing cluster separability more effectively in many settings.
• We embed counterfactual separation directly into the clustering objective, proposing **cf-means**, a separation-aware variant of k-means with reduced sensitivity to initialization.
• We extend this framework to deep clustering with **deep cf-means**, jointly optimizing reconstruction, compactness, and counterfactual separation to learn enhanced latent representations.

1. Introduction

• Clustering is a fundamental unsupervised learning task, widely used for data exploration, representation learning, and pattern discovery.
• Evaluating clustering quality and selecting the number of clusters remain challenging due to the absence of ground-truth labels.
• Existing internal indices mainly rely on pairwise distances or centroid dispersion, often failing to capture meaningful cluster separation.
• Counterfactual explanations provide a principled way to measure how close a data point lies to the decision boundary of competing clusters.
• We define **CFQ**, a **separation-aware clustering quality index** based on counterfactual distances.
• We develop two separation-aware clustering algorithms, **cf-means** and **deep cf-means**, which incorporate counterfactual-based separation directly into shallow and deep clustering objectives.

3.The cf-means and deep cf-means clustering algorithms

Algorithm 2 The cf-means algorithm

Require: Data matrix X, number of clusters K, initial centers $\{\mu_k\}$, initial weight λ_0, λ-decay γ, learning rate η, number of iterations T
Initialize $\lambda \leftarrow \lambda_0$
for $t = 1$ to T **do**
  (Assignment) For each point x_i, compute

$$c_i = \arg\min_k \|x_i - \mu_k\|.$$

  (Counterfactual Assignment) For each point x_i, compute the nearest alternative cluster

$$a_i = \arg\min_{k \neq c_i} \|x_i - \mu_k\|.$$

  (Counterfactual Gradient) Initialize $\nabla_{\mathrm{cf}}\mu_k = 0$ for all k. For each point x_i:

$$v_i = \mu_{a_i} - \mu_{c_i},$$
$$\delta_i = v_i^\top x_i - \tfrac{1}{2}(\|\mu_{a_i}\|^2 - \|\mu_{c_i}\|^2),$$

  and, if $\|v_i\|^2$ is not too small, update

$$\nabla_{\mathrm{cf}}\mu_{c_i} \mathrel{+}= \lambda \cdot \Delta_{c_i}(x_i), \qquad \nabla_{\mathrm{cf}}\mu_{a_i} \mathrel{+}= \lambda \ \ \Delta_{a_i}(x_i),$$

  where $\Delta_{c_i}(x_i) = \frac{\partial \ell_{\mathrm{cf}}(x_i)}{\partial \mu_{c_i}} = \frac{2\delta_i|v_i|^2(\mu_{c_i} - x_i) + 2\delta_i^2(\mu_{a_i} - \mu_{c_i})}{|v_i|^4}$,
  and $\Delta_{a_i}(x_i) = \frac{\partial \ell_{\mathrm{cf}}(x_i)}{\partial \mu_{a_i}} = \frac{2\delta_i|v_i|^2(x_i - \mu_{a_i}) - 2\delta_i^2(\mu_{a_i} - \mu_{c_i})}{|v_i|^4}$ denote the counterfactual gradient contributions for the assigned and alternative centers, respectively.
  (Center Update) For each cluster k:

$$g_k^{\mathrm{kmeans}} = 2 \cdot \mathrm{E}_{i:c_i=k}(\mu_k - x_i),$$
$$g_k = g_k^{\mathrm{kmeans}} - \nabla_{\mathrm{cf}}\mu_k,$$
$$\mu_k \leftarrow \mu_k - \eta\, g_k.$$

  (Weight Decay) $\lambda \leftarrow \gamma\lambda$
end for
(k-means refinement) run k-means initialized at centers μ_k
return final centers $\{\mu_k\}$ and assignments $\{c_i\}$

Algorithm 3 Deep cf-means clustering algorithm

Require: Dataset $\mathcal{D} = \{x_i\}_{i=1}^N$, autoencoder (f_ϕ, g_ψ), number of clusters K, reconstruction weight λ_{rec}, k-means weight λ_{k}, counterfactual weight λ_{cf}.

Pretraining Phase
for epoch $= 1 \ldots T_{\mathrm{pre}}$ **do**
  for minibatch $\mathcal{B} \subset \mathcal{D}$ **do**
    $z_i = f_\phi(x_i)$ for all $x_i \in \mathcal{B}$
    $\hat{x}_i \ : g_\psi(z_i)$
    Compute reconstruction loss:

$$\mathcal{L}_{\mathrm{rec}} = \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} \|x_i - \hat{x}_i\|^2$$

    Update ϕ, ψ via gradient descent on $\mathcal{L}_{\mathrm{rec}}$.
  end for
end for

Initialize Cluster Centers
Sample K points from the latent representations:

$$\mu_k \leftarrow f_\phi(x_{j_k}), \quad k = 1, \ldots, K$$

Joint Training Phase
for epoch $= 1 \ldots T_{\mathrm{joint}}$ **do**
  for minibatch $\mathcal{B} \subset \mathcal{D}$ **do**
    $z_i = f_\phi(x_i)$, $\hat{x}_i = g_\psi(z_i)$
    Compute nearest and second-nearest centers:

$$c_i = \arg\min_k \|z_i - \mu_k\|^2, \qquad a_i = \arg\min_{k \neq c_i} \|z_i - \mu_k\|^2$$

    Compute k-means loss:

$$\mathcal{L}_{\mathrm{k}} = \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} \|z_i - \mu_{c_i}\|^2$$

    Compute counterfactual separation loss:

$$v_i = \mu_{a_i} \quad \mu_{c_i}, \qquad b_i = \tfrac{1}{2}\left(\|\mu_{a_i}\|^2 \quad \|\mu_{c_i}\|^2\right)$$

$$\mathcal{L}_{\mathrm{cf}} = \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} \frac{\left(v_i^\top z_i - b_i\right)^2}{\|v_i\|^2}$$

    Compute total loss:

$$\mathcal{L}_{\mathrm{deep}} = \lambda_{\mathrm{rec}}\mathcal{L}_{\mathrm{rec}} + \lambda_{\mathrm{k}}\mathcal{L}_{\mathrm{k}} - \lambda_{\mathrm{cf}}\mathcal{L}_{\mathrm{cf}}$$

    Update autoencoder parameters (ϕ, ψ) and centers $\{\mu_k\}$ using gradient descent
  end for
end for

return Trained autoencoder (f_ϕ, g_ψ) and cluster centers $\{\mu_k\}$

2. Counterfactual-based Quality Score

• The distance $Sep(j, l)$ of cluster C_j to C_l is defined as the sum of squared counterfactual distances of the points of C_j to cluster C_l:

$$Sep(j, l) = \sum_{x_i \in C_j} d_{jl}^2(x_i).$$

•The separation $S(j)$ of cluster C_j is defined as its minimum separation to any other cluster:

$$S(j) = \min_{l,\, l \neq j} Sep(j, l).$$

•The total separation $TS(C)$ of the clustering solution is computed as:

$$TS(C) = \sum_{j=1}^{k} S(j).$$

•The total variance $TV(C)$ of the clustering solution is defined as:
• Our clustering quality measure (CFQ) is the ratio of the two terms:

$$CFQ(C) = \frac{TS(C)}{TV(C)}.$$

4. Experiments

Table 2: NMI statistics over 100 trials for each dataset.

Dataset	Method	Mean	Std
Optdigits	CFQ	0.711	0.118
	SIL	0.725	0.121
	VRC	0.477	0.204
Pendigits	CFQ	0.641	0.135
	SIL	0.629	0.149
	VRC	0.543	0.178
MNIST Embeddings	CFQ	0.712	0.138
	SIL	0.691	0.174
	VRC	0.542	0.194
USPS	CFQ	0.561	0.149
	SIL	0.465	0.207
	VRC	0.394	0.179

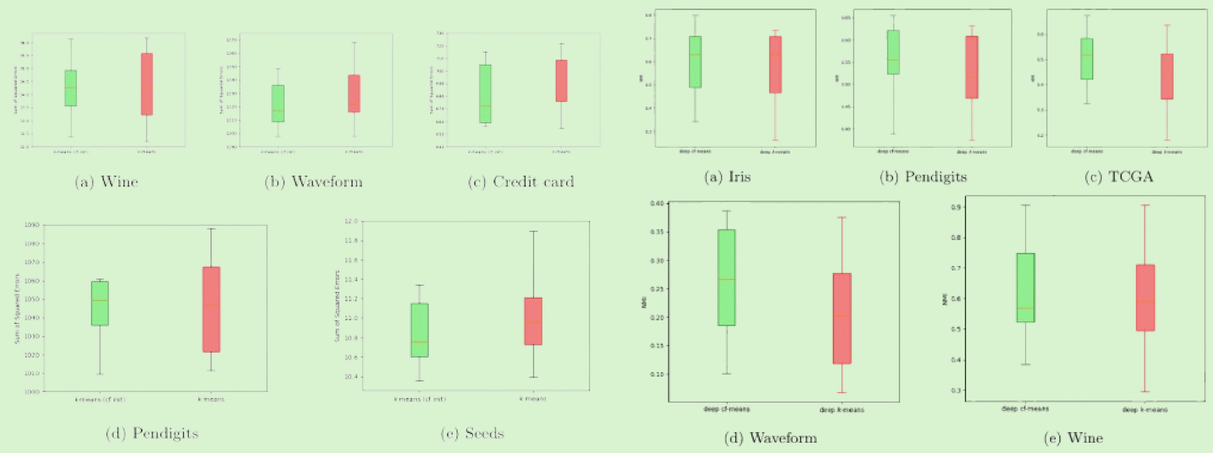

Figure 23: Comparison of the k-means clustering error across 50 runs for five datasets, using cf-means and standard k-means, both starting from the same random initialization. Each subfigure presents SSE box plots of the two methods for a specific dataset (lower is better). Across most datasets, cf-means yields lower and more stable final errors, demonstrating its effectiveness in attaining better local minima that k-means.

Figure 24: Box plots of NMI scores for the compared deep clustering methods (deep cf-means and deep k-means) across five benchmark datasets. Each plot summarizes the NMI distribution for each dataset over ten independent runs using fro both methods (i) the same initial pretrained AE and (ii) identical random latent center initializations.

5. Conclusions

• We introduced counterfactual distances as an effective way to quantify cluster separation.
• We proposed CFQ, a separation-aware clustering quality index that complements existing internal validation measures.
• We embedded counterfactual-based separation directly into clustering objectives through **cf-means**, improving robustness.
• We extended this framework to representation learning with **deep cf-means**, yielding more compact and separated latent clusters.
• Overall, counterfactual-based separation provides an approach for both clustering evaluation and separation-aware clustering algorithms.

Counterfactuals for
Clustering and Deep Clustering:
Explainability, Fairness and Quality
Project Number: 15940
Hellenic Foundation for Research & Innovation (H.F.R.I)
“Basic Research Funding (Horizontal Support for all Sciences)”
National Recovery and Resilience