The Density Awakens: Boosting-Enhanced Clustering

Abhishek Sharma

April 4, 2025

Abstract

This paper presents a novel hybrid approach to clustering by incorporating a supervised proxy for local density into traditional KMeans. Our technique uses regionally normalized Local Outlier Factor (LOF) scores to define a density proxy, trained via Gradient Boosting Regressors (GBR). We augment the feature space with predicted density scores and evaluate clustering performance through both standard and cohesion-aware metrics such as Silhouette Score, Davies–Bouldin Index, Calinski–Harabasz Index, and average Gower distance. This technique improves the interpretability and robustness of KMeans without altering its core algorithmic simplicity.

1 Introduction

Clustering high-dimensional data remains a central challenge in unsupervised learning. Traditional methods such as KMeans perform well for globular clusters but fail in data with varying densities or outliers. In this paper, we introduce a Gradient Boosted Density-Aware Clustering method that augments the KMeans feature space using a learned proxy for local density.

Our proposed framework addresses key limitations in classical clustering by modeling density explicitly using supervised learning. We employ Local Outlier Factor (LOF) as a density indicator and generalize its behavior with a Gradient Boosted Regressor (GBR). By feeding this density proxy back into KMeans, the algorithm becomes capable of respecting local data structure, resulting in more meaningful cluster formations.

The overall methodology retains the scalability and speed of KMeans while infusing it with density-awareness, typically associated with methods like DBSCAN. This fusion enables interpretable and robust clustering applicable to complex, real-world data.

Why This Approach?

KMeans lacks the notion of point density and is easily swayed by outliers and varying densities whereas unlike global measures, LOF focuses on comparing the density around a point to that of its neighbors and captures micro-level density variations, ideal for spotting subtle structure. Our pipeline works on:

- Learning an interpretable density function through supervised regression
- Leveraging GBR for its ability to model non-linear interactions and provide SHAP based explainability
- Enhancing the geometric representation of features via density-based augmentation

This novel intersection of unsupervised local structure and supervised learning empowers KMeans to operate in density-variant spaces while remaining computationally efficient.

2 Motivation and Rationale

Most clustering algorithms either operate purely geometrically (e.g., KMeans) or assume knowledge of density structure (e.g., DBSCAN). However, density-aware clustering methods like DBSCAN suffer from sensitivity to hyperparameters and poor scalability. On the other hand, KMeans lacks sensitivity to local density variations, often splitting dense regions or merging sparse ones.

Our approach introduces a third paradigm: **learning density** via proxy modeling. We generate a differentiable and smooth surrogate for local density using a supervised learning technique (Gradient Boosting Regressors) trained on Local Outlier Factor (LOF) scores. This proxy captures local structural information in a generalizable form, allowing KMeans to incorporate density-awareness without modifying its algorithm.

The advantages of this design include:

- **Scalability**: We retain the simplicity and computational efficiency of KMeans.
- **Interpretability**: SHAP values from GBR allow feature-wise explanation of density predictions.
- **Robustness**: Region-wise normalization smoothens LOF and reduces noise.
- **Hybrid Evaluation**: Combining Silhouette, Davies-Bouldin, Calinski-Harabasz, and Gower distance gives both geometric and semantic cohesion evaluation.

Why Region-wise LOF Normalization?

LOF is inherently local and sensitive to neighborhood size and global outliers. Instead of applying LOF globally, we normalize it within mini-regions defined by a coarse-grained KMeans partition. This:

- Maintains locality and improves robustness to noise/outliers
- Allows smoother, bounded interpretation of LOF as a continuous density score
- Enables region-aware scaling that adjusts to data heterogeneity

Why Gradient Boosted Regressor (GBR)?

Gradient Boosting is an ensemble of decision trees that captures complex, non-linear interactions. We use it to learn a mapping from original features to LOF-based density, which provides:

- Differentiable approximation of an otherwise non-parametric measure
- Scalability to larger datasets
- Explainability via SHAP analysis

This turns LOF from a black-box outlier detector into a reusable, interpretable feature transformer.

Why Feature Augmentation?

By appending the predicted density to the original features, we steer KMeans toward respecting cluster density variations. This:

- Biases centroid movement to consider dense vs. sparse areas
- Simulates density-aware separation without switching to DBSCAN
- Enhances separability and improves intra-cluster compactness

Potential Improvements

- Use of contrastive learning to generate structure-aware embeddings for density estimation
- Multiscale normalization for combining global and local LOF
- Alternative regressors such as LightGBM for speed improvements
- Use of semi-supervised fine-tuning to adapt clusters post-hoc

3 Theoretical Formulation and Related Work

Local Outlier Factor (LOF) as Density Proxy

Given a data point $\mathbf{x}_i$, its Local Outlier Factor (LOF) is defined as:

$$\text{LOF}_k(\mathbf{x}_i) = \frac{1}{|N_k(\mathbf{x}_i)|} \sum_{\mathbf{x}_j \in N_k(\mathbf{x}_i)} \frac{\text{lrd}_k(\mathbf{x}_j)}{\text{lrd}_k(\mathbf{x}_i)}$$

where $N_k(\mathbf{x}_i)$ denotes the k-nearest neighbors and $\text{lrd}_k(\mathbf{x})$ is the local reachability density. This value provides an inverse measure of how "dense" or "central" a point is in its local neighborhood.

Region-wise Normalization

Rather than using the raw LOF scores globally, we divide the dataset into R regions using MiniBatchKMeans and normalize the LOF scores within each region r as:

$$\tilde{LOF}_r(\mathbf{x}_i) = 1 - \frac{LOF(\mathbf{x}_i) - \min_r}{\max_r - \min_r + \epsilon}$$

where $\min_r$ and $\max_r$ are the minimum and maximum LOF values in region r.

Supervised Proxy Learning via GBR

Let $f : \mathbb{R}^d \to \mathbb{R}$ be the regressor that approximates the normalized LOF using input features $\mathbf{x}_i$:

$$\hat{y}_i = f(\mathbf{x}_i) \approx \tilde{LOF}(\mathbf{x}_i)$$

This is learned by minimizing the mean squared error (MSE) between predictions and target LOF proxies.

Feature Space Augmentation

The final feature space for clustering becomes:

$$\mathbf{x}'_i = [\mathbf{x}_i, \hat{y}_i]$$

This extended representation biases clustering toward regions with consistent density estimates.

Related Work

- LOF was introduced by Breunig et al. (2000) as a method for unsupervised outlier detection.
- Gradient Boosting Machines, especially Friedman's implementation (2001), are powerful for capturing complex non-linear patterns.
- Previous hybrid clustering methods include semi-supervised KMeans and distance-weighted variants, but few combine GBDT-based density estimation.

4 Methodology

4.1 Step 1: Synthetic Data Generation

We simulate a dataset of 10,000 samples, with 5 highly correlated and 5 low-correlated features. An additional 50 outliers are added to represent atypical observations. Features are standardized for downstream analysis.

4.2 Step 2: Density Proxy via Region-Wise LOF

LOF provides a local density estimation based on a point's neighborhood. To improve interpretability and robustness, we compute LOF within local mini-regions defined by KMeans clustering:

$$\mathrm{LOF}_k(\mathbf{x}_i) = \frac{1}{|N_k(\mathbf{x}_i)|} \sum_{\mathbf{x}_j \in N_k(\mathbf{x}_i)} \frac{\mathrm{lrd}_k(\mathbf{x}_j)}{\mathrm{lrd}_k(\mathbf{x}_i)} \tag{1}$$

where $N_k(\mathbf{x}_i)$ is the set of k-nearest neighbors of point $\mathbf{x}_i$ and $\mathrm{lrd}_k(\cdot)$ is the local reachability density.
Each region is normalized independently:

$$\text{Normalized Density}_i = 1 - \frac{LOF_i - LOF_{\min}}{LOF_{\max} - LOF_{\min} + \epsilon} \tag{2}$$

This ensures comparability across regions.

4.3 Step 3: Proxy Regression using GBR

We fit a Gradient Boosting Regressor to learn the LOF proxy. To prevent overfitting and enhance generalization, we apply early stopping using a validation split.

The learned function $f_{\mathrm{GBR}}(\mathbf{x}) \approx \mathrm{LOF}_{\mathrm{norm}}(\mathbf{x})$ now generalizes LOF behavior to unseen data in a scalable manner. This bridges unsupervised local behavior and supervised prediction.

4.4 Step 4: Feature Augmentation and Clustering

We augment the original feature space with the predicted density and apply KMeans clustering. This density-aware feature enhances separability in cases of variable cluster density.

$$X_{\mathrm{aug}} = [X \mid f_{\mathrm{GBR}}(X)] \tag{3}$$

This process changes the geometry of the feature space in a meaningful way:

In denser regions, the appended density score is closer to 1, influencing KMeans centroids to shift toward denser areas. In sparser regions, the score is lower, discouraging KMeans from aggregating outliers or noise.

This augmentation serves multiple purposes:

- It enables KMeans, which normally operates only on spatial distances, to also consider density structure without modifying its algorithm.
- It reduces false splits in dense areas and false merges in sparse ones.
- It provides a bridge between unsupervised learning (LOF) and supervised approximation (GBR), resulting in a more robust and interpretable clustering outcome.

4.5 Step 5: Evaluation Metrics

We compare vanilla and augmented KMeans using:

- Silhouette Score (higher is better)
- Davies–Bouldin Index (lower is better)
- Calinski–Harabasz Index (higher is better)
- Gower Distance (lower is better)

4.6 Step 6: Visual and Quantitative Analysis

PCA and UMAP are used to visualize the clustering. SHAP values are computed to understand which features contribute most to the density proxy.

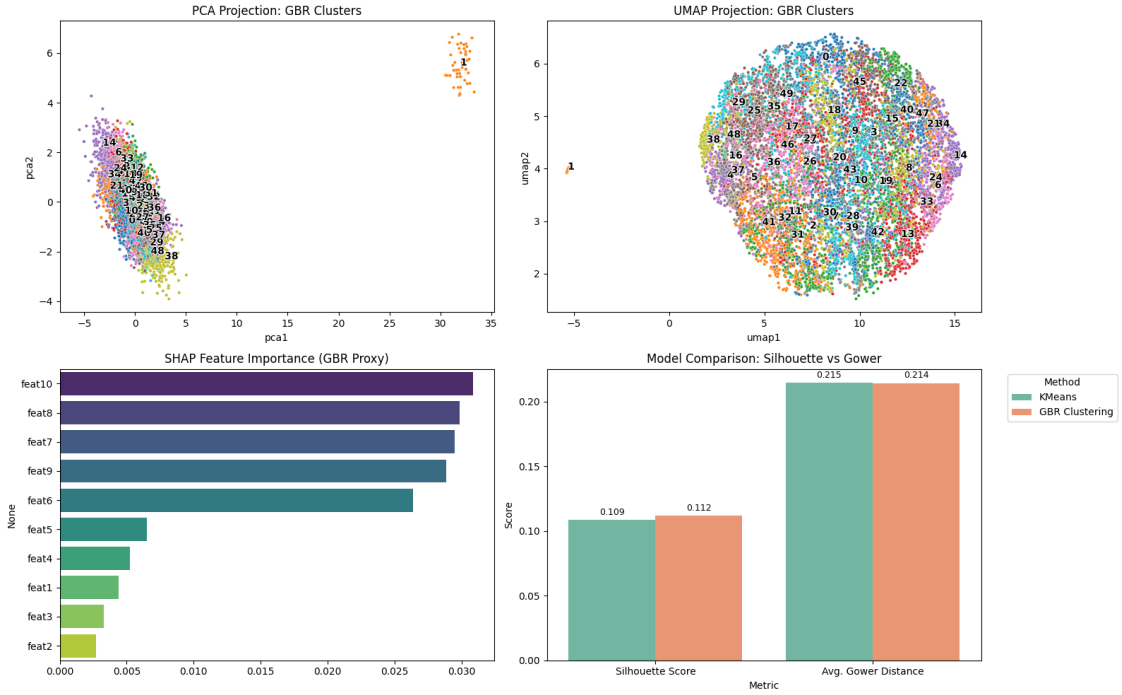

Figure 1: Gradient-Boosted Density-Aware Clusters & Evaluation Metrics

5 Novel Contributions

- **Proxy Learning with GBDTs:** KMeans is enhanced with a learned density signal, approximating DBSCAN-like behavior using regression.
- **Region-wise Normalized LOF:** Adapts LOF scores locally to ensure smooth proxy estimation and consistency across regions.

6 Conclusion

This method provides a lightweight yet effective alternative to traditional density-based clustering. By learning density proxies and feeding them to KMeans, we bridge the interpretability of density methods with the scalability of centroid-based clustering. This methodology enables KMeans to respect both geometry and local structure without compromising computational efficiency.

7 References

- Breunig, M. M., Kriegel, H. P., Ng, R. T., & Sander, J. (2000). LOF: Identifying density-based local outliers. ACM sigmod record.
- Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of statistics.
- Kaufman, L., & Rousseeuw, P. J. (2009). Finding groups in data: an introduction to cluster analysis. Wiley.