Decision-Guided Clustering for Energy-Efficient LoRaWAN Networks Using Data Representations

Cristofe Coelho Lopes da Rocha
Centro de Informática (CIn)
Universidade Federal de Pernambuco
Recife, Brazil, 50740-560
cclr@cin.ufpe.br

Paulo Filipe Cândido Barbosa
Campus Petrolina
Universidade de Pernambuco
Petrolina, Brazil, 56328-900
paulo.fcbarbosa@upe.br

Renato Mariz de Moraes
Centro de Informática (CIn)
Universidade Federal de Pernambuco
Recife, Brazil, 50740-560
renatomdm@cin.ufpe.br

***Abstract*—Energy efficiency and scalability in LoRaWAN are challenged by heterogeneous communication conditions and clustering strategies that operate directly on the original feature space. This paper proposes a decision-guided clustering approach in which Random Forest (RF) class probabilities provide a learned representation for K-Means and DBSCAN. Experiments with LoRaSimPlus show that K-Means particularly benefits from this representation: the Silhouette score increases from 0.559 to 0.836, the Davies–Bouldin index decreases from 0.476 to 0.303, mean transmission energy is reduced by 43.4%, and the coefficient of variation decreases from 0.634 to 0.273. In contrast, DBSCAN reduces mean transmission energy and improves within-cluster energy homogeneity but does not achieve the same geometric gains, showing that the effectiveness of learned representations depends on the clustering paradigm. Overall, RF probability representations can improve both cluster structure and energy-related behavior in heterogeneous LoRaWAN networks, particularly when combined with centroid-based clustering.**

***Index Terms*—decision-guided clustering, energy-efficient clustering, IoT, LoRaWAN, machine learning, random forest.**

I. Introduction

Wireless Sensor Networks (WSNs) and the Internet of Things (IoT) support a wide range of applications, including smart cities, precision agriculture, industrial automation, environmental monitoring, and healthcare. Their continued expansion has increased demand for communication technologies that connect large numbers of battery-powered devices while providing long-range connectivity, scalability, reliability, and energy-efficient operation [1].

Among Low-Power Wide-Area Network (LPWAN) technologies, LoRaWAN has become widely adopted for its long communication range, low deployment costs, and reduced energy requirements. Nevertheless, increasing network scale and heterogeneity can intensify channel contention, packet collisions, communication variability, and energy depletion [2]–[4].

Clustering provides a mechanism for organizing heterogeneous network observations and supporting energy-aware resource management. However, conventional clustering algorithms, such as K-Means and DBSCAN, generally operate directly on the original feature space. In large-scale IoT datasets, this space may contain substantial redundancy and overlapping information with energy-consumption regimes, making relevant structures difficult to distinguish [5].

In this context, supervised learning can provide additional information to guide subsequent clustering. Rather than being used exclusively for prediction, a supervised model can transform labeled observations into a representation that reflects the decision structure learned during training [6], [7]. In this work, Random Forest (RF) is used to map network observations into class-probability vectors. These probability vectors provide a compact, decision-informed representation in which the relationships among observations reflect the classification patterns learned by the RF model. K-Means and DBSCAN are subsequently applied to this transformed space instead of operating exclusively on the original network features.

Motivated by this perspective, this work investigates Decision-Guided Clustering for Energy-Efficient LoRaWAN Networks using RF-derived data representations. Network data are generated through LoRaSimPlus simulations and used to train the RF model [8]. The resulting class-probability representation is subsequently used by K-Means and DBSCAN. The central hypothesis is that incorporating supervised decision information before clustering provides a more informative organization of heterogeneous LoRaWAN observations than clustering directly in the original feature space.

The rest of this paper is organized as follows. Section II reviews related work and identifies the research gap. Section III presents the proposed method and experimental setup. Section IV presents and discusses the results, and Section V concludes the paper.

II. Related Work and Research Gap

LoRaWAN relies on a shared wireless medium and ALOHA-based channel access, making dense deployments susceptible to collisions, interference, and energy inefficiency [2], [3], [9]. Clustering has therefore been investigated to organize devices with similar communication characteristics and improve resource utilization [10]–[12].

Conventional methods, such as K-Means and DBSCAN, typically operate directly on the original feature space [5], [13]. However, heterogeneous LoRaWAN observations may exhibit nonlinear relationships among transmission parameters, link conditions, and energy consumption that are not adequately captured by geometric proximity alone.

979-8-3195-0167-7/26/$31.00 ©2026 IEEE

Machine-learning techniques have increasingly been applied to LoRaWAN optimization [14]–[16]. Random Forest (RF), in particular, can capture nonlinear relationships while providing class-membership probabilities that encode the model's learned decision information. These probabilities can therefore define an alternative representation space for subsequent clustering.

Despite advances in clustering and learning-based LoRaWAN optimization, supervised information is still primarily used for prediction rather than as a representation for unsupervised clustering. To address this gap, we use RF class-probability outputs as a decision-informed representation and apply K-Means and DBSCAN to the resulting probability space. The approach is evaluated using LoRaWAN data generated with LoRaSimPlus [8].

III. Proposed Method and Experimental Setup

The proposed decision-guided clustering approach combines supervised learning and unsupervised clustering to organize LoRaWAN observations according to communication and energy-related behavior. Instead of clustering exclusively in the original feature space, a Random Forest (RF) model transforms each observation into a class-probability representation, which is subsequently used by K-Means and DBSCAN.

Fig. 1 illustrates the overall architecture of the proposed approach and its comparison with the baseline. Both paths start from the same preprocessed LoRaSimPlus data. In the baseline, K-Means and DBSCAN operate directly on the original feature space X, whereas the proposed approach first uses the RF to transform each observation into a class-probability representation P through `predict_proba`. The same clustering algorithms are then applied to this decision-informed probability space, allowing the effect of the learned representation to be evaluated independently of the clustering algorithm itself.

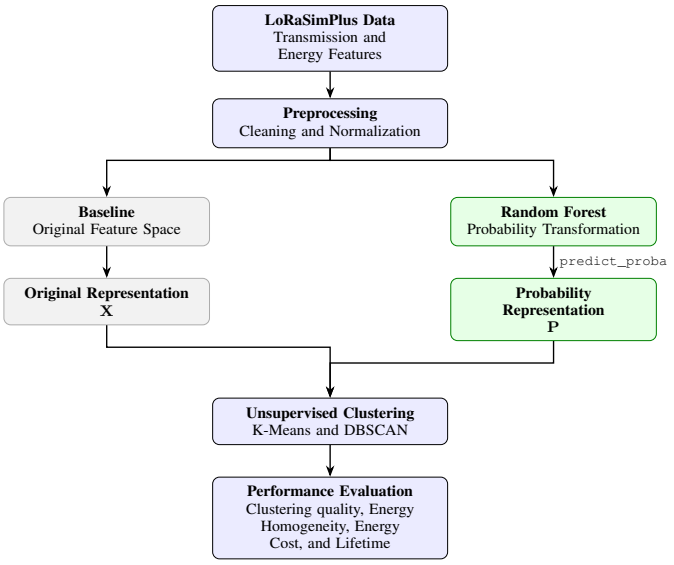


Fig. 1. Decision-guided clustering architecture. The baseline applies K-Means and DBSCAN to the original feature space, whereas the proposed approach transforms LoRaWAN observations into a Random Forest class-probability space before clustering.

The approach comprises three stages: (i) LoRaWAN data generation and preprocessing, (ii) RF-based probability-space transformation, and (iii) decision-guided clustering.

A. LoRaSimPlus Dataset Generation

The experimental dataset was generated using **LoRaSimPlus**, which models LoRaWAN packet transmissions, propagation, collisions, radio parameters, and energy consumption [8]. The simulated network follows the LoRaWAN star-of-stars architecture, with battery-powered Class A end devices communicating with a centralized gateway.

The number of devices ranges from 500 to 1000 within a $2000 \times 2000 \;\; m^2$ deployment area. Nodes follow stochastic hotspot distributions to represent heterogeneous IoT deployments and localized communication contention. The simulations use SF7–SF12 spreading factors, 125 kHz bandwidth, coding rate 4/5, 14 dBm transmission power, and 20-byte payloads. A total of 348 independent Monte Carlo executions were performed. Table I summarizes the main parameters.

TABLE I
Main LoRaSimPlus simulation parameters [8].

Parameter	Value
Deployment area	$2000 \times 2000\ m^2$
Gateway	1 central gateway
Sensor nodes	500–1000
Distribution	Stochastic hotspots
Device class	LoRaWAN Class A
Frequency	915 MHz
Bandwidth	125 kHz
Spreading factor	SF7–SF12
Coding rate	4/5
TX power	14 dBm
Payload	20 Bytes
Executions	348 Monte Carlo runs

Communication logs from all executions were merged and preprocessed by removing invalid and incomplete observations and normalizing numerical attributes. Transmission energy was discretized into five classes using quintile-based thresholds (`q=5`), resulting in five energy-consumption regimes with approximately equal numbers of observations. This quantile-based discretization was adopted to represent different energy-consumption levels while reducing class imbalance during supervised learning. The RF input vector is defined as

$$\mathbf{x}_i = [\texttt{sf}, \texttt{toa_ms}, \texttt{bw_hz}, \texttt{code_rate}], \qquad (1)$$

where `sf`, `toa_ms`, `bw_hz`, and `code_rate` represent the spreading factor, Time-on-Air, bandwidth, and coding rate, respectively. The dataset was divided into training and test sets using a stratified 70%/30% split.

B. RF-Based Probability Representation

The RF model is used not only as a classifier but also as a transformation mechanism. For each observation $\mathbf{x}_i$, the trained model produces

$$\mathbf{p}_i = \texttt{predict_proba}(\mathbf{x}_i), \qquad \mathbf{p}_i \in \mathbb{R}^5, \qquad (2)$$

where each component represents the estimated membership probability for one of the five energy classes employed. Thus, the transformation

$$\mathbf{x}_i \xrightarrow{\text{RF}} \mathbf{p}_i \qquad (3)$$

maps the original transmission features into a compact decision-informed probability space. Observations with similar energy-related behavior can therefore, become closer in the transformed space even when their original feature values differ.

RF hyperparameters were selected using five-fold stratified cross-validation and GridSearchCV with macro F1-score as the optimization criterion. The selected configuration used 100 trees, unrestricted maximum depth, `min_samples_split` $= 10$, and `min_samples_leaf` $= 4$. The resulting class-probability outputs are subsequently used as a decision-informed representation for downstream clustering, as indicated in Fig. 1.

C. Decision-Guided Clustering

To isolate the effect of the learned representation, K-Means and DBSCAN are evaluated both in the standardized original feature space (baseline) and in the RF probability space (proposed approach). Therefore, four configurations are compared: K-Means, DBSCAN, K-Means (RF), and DBSCAN (RF).

For K-Means, $k = 5$ is adopted to match the five energy regimes. For DBSCAN, the RF-space configuration uses $\varepsilon = 0.1$ and $\mathrm{MinPts} = 5$. No modifications are introduced to the clustering algorithms themselves; the main distinction lies in the representation provided as their input.

The proposed pipeline is summarized as

$$\text{LoRaSimPlus} \rightarrow \mathbf{X} \rightarrow \text{RF} \rightarrow \mathbf{P} \rightarrow \{\text{K-Means}, \text{DBSCAN}\}. \quad (4)$$

Clustering performance is assessed using geometric and energy-aware metrics, while network sustainability is evaluated through mean transmission energy consumption and estimated lifetime. Mean energy, reported in mJ, represents the average energy consumption associated with each clustering configuration. Energy homogeneity is quantified using the coefficient of variation (CV), which measures the relative dispersion of transmission energy consumption within each cluster. For each cluster c, the CV is computed as

$$\mathrm{CV}_c = \frac{\sigma_{E,c}}{\mu_{E,c}}, \quad (5)$$

where $\sigma_{E,c}$ and $\mu_{E,c}$ denote the standard deviation and mean transmission energy consumption within cluster c, respectively. The reported CV is the average of the cluster-level CV values, with lower values indicating greater within-cluster energy homogeneity. This comparison determines whether the RF-derived probability representation provides a more informative clustering space than the original LoRaWAN transmission features.

IV. Results and Discussion

The RF classifier achieved an accuracy of 0.7844 (95% confidence interval: [0.7808, 0.7881]) and a macro F1-score of 0.7095. Although classification performance is not the primary objective of the proposed framework, these results indicate that the RF captures useful discriminatory information among the considered energy regimes. More importantly, the proposed approach does not rely solely on predicted class labels; instead, it uses continuous class-probability vectors that preserve similarities and predictive uncertainty among energy regimes. These probability outputs define the representation subsequently provided to the clustering algorithms. Therefore, the supervised stage is evaluated not only by its classification performance, but primarily by the utility of the learned representation for downstream clustering.

TABLE II
Baseline and RF-guided clustering performance.

Method	Sil.↑	DB↓	Energy (mJ)↓	CV↓	Lifetime (year)↑
K-Means	0.559	0.476	49.137	0.634	198.9
K-Means (RF)	**0.836**	**0.303**	27.811	0.273	273.9
DBSCAN	0.581	0.355	45.980	0.527	**377.3**
DBSCAN (RF)	0.351	0.418	**23.545**	**0.212**	245.3

Table II compares K-Means and DBSCAN in the original and RF probability spaces, isolating the effect of the learned representation while keeping the clustering algorithms unchanged.

For K-Means, the RF representation substantially improves the geometric structure of the clusters. The Silhouette score increases from 0.559 to 0.836, corresponding to a 49.6% improvement, while the Davies–Bouldin index decreases from 0.476 to 0.303, a reduction of 36.3%. These metrics provide complementary evidence: the higher Silhouette score indicates greater intra-cluster cohesion and inter-cluster separation, whereas the lower Davies–Bouldin index indicates better separation among neighboring clusters.

The energy-related results also improve substantially for K-Means. Mean transmission energy consumption decreases from 49.137 mJ to 27.811 mJ, corresponding to a reduction of approximately 43.4%, while the CV decreases from 0.634 to 0.273, a reduction of 56.9%, indicating substantially greater within-cluster energy homogeneity. Moreover, the estimated lifetime increases from 198.9 to 273.9 years, corresponding to an improvement of approximately 37.7%. Taken together, these results indicate that the RF probability representation is particularly compatible with the centroid-based partitioning mechanism of K-Means, improving cluster separation and within-cluster energy homogeneity while reducing mean energy and increasing estimated lifetime.

A different behavior is observed for DBSCAN. In the RF probability space, the Silhouette score decreases from 0.581 to 0.351, while the Davies–Bouldin index increases from 0.355 to 0.418. Thus, unlike K-Means, DBSCAN does not obtain geometric improvements from the transformed representation. This result is consistent with the different mechanisms of the two algorithms. K-Means relies on distances to centroids, whereas DBSCAN identifies clusters according to local density and neighborhood connectivity. Transforming the observations into class-probability vectors changes the geometry and local density structure of the input space and can therefore alter the neighborhoods identified using fixed ε and MinPts parameters.

Despite the weaker geometric results, DBSCAN (RF) exhibits substantial improvements in the energy-related metrics. Mean energy decreases from 45.980 mJ to 23.545 mJ, corresponding to a reduction of approximately 48.8%, while the CV

decreases from 0.527 to 0.212, a reduction of approximately 59.8%. The latter is the lowest CV among all evaluated configurations, indicating the highest within-cluster energy homogeneity. However, these improvements do not extend to the estimated lifetime: the original DBSCAN configuration achieves the highest lifetime, 377.3 years, compared with 245.3 years for DBSCAN (RF). This result highlights that improvements in mean energy consumption and within-cluster energy homogeneity do not necessarily translate into longer estimated network lifetime.

The combined results show that the effectiveness of the learned representation depends on both the clustering algorithm and the evaluation criterion. For K-Means, the RF probability space provides consistent improvements in geometric quality, mean energy, within-cluster energy homogeneity, and estimated lifetime. For DBSCAN, the same representation reduces mean energy and improves within-cluster energy homogeneity, but at the cost of poorer geometric quality and lower estimated lifetime. Thus, the learned representation and the clustering paradigm should be considered jointly when evaluating decision-guided clustering for LoRaWAN networks.

Overall, the results support the central premise of the proposed approach: supervised class-probability information can provide a useful representation for subsequent unsupervised clustering without modifying the clustering algorithms themselves. The benefits are particularly consistent for K-Means, where improvements are observed across both geometric and energy-related metrics. For DBSCAN, the results reveal a trade-off between within-cluster energy homogeneity and geometric structure, suggesting that density-based clustering may require parameter adaptation after transformation to the RF probability space.

V. Conclusion

This work investigated decision-guided clustering for energy-efficient LoRaWAN networks using Random Forest class probabilities as a learned representation for K-Means and DBSCAN. Rather than modifying the clustering algorithms, the proposed approach incorporates supervised energy-related information by transforming the original feature space into a decision-informed probability space prior to clustering.

Results show that K-Means particularly benefits from the RF representation: the Silhouette score increased by 49.6%, the Davies–Bouldin index decreased by 36.3%, mean transmission energy decreased by 43.4%, and CV decreased by 56.9%, while the estimated lifetime increased by 37.7%. For DBSCAN, the RF representation reduced the mean energy by 48.8% and achieved the lowest CV (0.212), indicating the highest within-cluster energy homogeneity among the evaluated configurations, but resulted in lower geometric quality and estimated lifetime. These findings indicate that the benefits of learned representations depend on the clustering paradigm, and that improvements in within-cluster energy homogeneity do not necessarily translate into better geometric structure or longer network lifetime.

Overall, RF probability representations provide an effective bridge between supervised learning and unsupervised network organization, with particularly consistent benefits for centroid-based clustering. The results also highlight the importance of jointly considering the learned representation and the clustering algorithm when designing energy-aware LoRaWAN solutions. Future work will investigate adaptive parameter tuning for density-based clustering and joint optimization of cluster quality, within-cluster energy homogeneity, and network lifetime under different LoRaWAN operating conditions.

Acknowledgment

This work was supported in part by Centro de Informática (CIn-UFPE).

References

[1] A. Al-Fuqaha, M. Guizani, M. Mohammadi, M. Aledhari, and M. Ayyash, "Internet of things: A survey on enabling technologies, protocols, and applications," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 4, pp. 2347–2376, 2015.

[2] M. Centenaro, L. Vangelista, A. Zanella, and M. Zorzi, "Long-range communications in unlicensed bands: The rising stars in the iot and smart city scenarios," *IEEE Wireless Communications*, vol. 23, no. 5, pp. 60–67, 2016.

[3] F. Adelantado, X. Vilajosana, P. Tuset-Peiro, B. Martinez, J. Melia-Segui, and T. Watteyne, "Understanding the limits of lorawan," *IEEE Communications Magazine*, vol. 55, no. 9, pp. 34–40, 2017.

[4] V. Bonilla, B. Campoverde, and S. G. Yoo, "A systematic literature review of lorawan: Sensors and applications," *Sensors*, vol. 23, no. 20, p. 8440, 2023.

[5] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD)*. AAAI Press, 1996, pp. 226–231.

[6] T. Finley and T. Joachims, "Supervised clustering with support vector machines," in *Proceedings of the 22nd International Conference on Machine Learning*, 2005, pp. 217–224.

[7] A. Ben-Hur, D. Horn, H. T. Siegelmann, and V. Vapnik, "Support vector clustering," *Journal of Machine Learning Research*, vol. 2, pp. 125–137, 2001.

[8] R. Wang, "Lorasimplus: Lorawan simulator developed based on simpy," https://github.com/Rich-King395/LoRaSimPlus, 2024, accessed: 2026-09-03.

[9] M. Bor, U. Roedig, T. Voigt, and J. M. Alonso, "Do lora low-power wide-area networks scale?" in *Proceedings of the 19th ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems (MSWiM)*, 2016, pp. 59–67.

[10] A. Alenezi, R. Djemame, and I. Ghafir, "Unsupervised learning clustering and dynamic transmission scheduling for efficient dense lorawan networks," *IEEE Access*, vol. 8, pp. 10 110–10 123, 2020.

[11] M. A. Lodhi, M. I. Ashraf, S. A. Chaudhry, M. Usama, and W. Yu, "Ai-enhanced resource allocation for lpwan-based lorawan: A hybrid tinyml and deep learning approach," *IEEE Access*, vol. 13, pp. 5421–5438, 2025.

[12] G. M. E. Rahman and K. A. Wahid, "Ldca: Lightweight dynamic clustering algorithm for iot-connected wide-area wsn and mobile data sink using lora," *IEEE Internet of Things Journal*, vol. 9, no. 2, pp. 1313–1326, 2022.

[13] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1. Berkeley, California: University of California Press, 1967, pp. 281–297.

[14] K. Alkhayyal and M. Mostafa, "Artificial intelligence-enabled energy efficiency techniques in lorawan networks: A survey," *Sensors*, vol. 24, no. 4, p. 4482, 2024.

[15] A. Farhad and J.-Y. Pyun, "Lorawan meets machine learning: A survey on enhancing performance," *Sensors*, vol. 23, no. 15, p. 6851, 2023.

[16] M. Alkhayyal and A. Mostafa, "Recent developments in ai and ml for iot: A systematic literature review on lorawan energy efficiency," *Sensors*, vol. 24, no. 14, p. 4482, 2024.