FA-Seed: Flexible and Active Learning-Based Seed Selection

Dinh-Minh Vu and Thanh-Son Nguyen *

Posted Date: 17 September 2025

doi: 10.20944/preprints202509.1405.v1

Keywords: fuzzy min–max neural network; fuzzy hyperbox; semi-supervised clustering; seed selection; active learning; label propagation; neuro-fuzzy systems; prototype-based clustering

 doi:10.20944/preprints202509.1405.v1

Article

FA-Seed: Flexible and Active Learning-Based Seed Selection

Dinh-Minh Vu [1], Thanh-Son Nguyen [2,*]

[1] Hanoi University of Industry, No. 298 Cau Dien Street, Tay Tuu Ward, Hanoi, Vietnam
[2] Academy of Finance, No. 58, Le Van Hien Street, Dong Ngac Ward, Hanoi, Vietnam
* Correspondence: sonnt@hvtc.edu.vn; Tel.:+84-0912-532-175

Abstract

This paper addresses the fundamental problem of seed selection in semi-supervised clustering, where the quality of initial seeds has a significant impact on clustering performance and stability. Existing methods often rely on randomly or heuristically selected seeds, which can propagate errors and increase dependence on expert labeling. To address these limitations, we propose FA-Seed, a flexible and adaptive model that integrates active querying with self-guided generalization within the framework of fuzzy hyperboxes. FA-Seed partitions the data into hyperboxes, evaluates seed reliability through measures of membership and association density, and propagates labels with an emphasis on label purity. The model demonstrates strong adaptability to complex and ambiguous data distributions in which cluster boundaries are vague or overlapping. The main contributions of FA-Seed include: (1) automatic estimation and selection of candidate instances for auxiliary supervision, (2) dynamic cluster expansion without retraining, (3) automatic detection and identification of structurally complex regions based on cluster characteristics, and (4) the ability to capture intrinsic cluster structures even when clusters vary in density and shape. Empirical evaluations on benchmark datasets, specifically the UCI and Computer Science collections, show that our approach consistently outperforms several state-of-the-art semi-supervised clustering methods.

Keywords: fuzzy min–max neural network; fuzzy hyperbox; semi-supervised clustering; seed selection; active learning; label propagation; neuro-fuzzy systems; prototype-based clustering

1. Introduction

Semi-supervised clustering (SSC) has emerged as an important paradigm in modern machine learning, as it leverages auxiliary information to improve clustering quality [1,2]. SSC enables learning from both limited labeled data and abundant unlabeled data, and is particularly useful in scenarios involving noisy or complexly distributed data. Pedrycz [3] demonstrated that, for practical applications which often require multiple intermediate ways of discovering structure in a dataset, performance can be significantly improved by using prior information; even a small proportion of labeled samples can markedly enhance clustering results.

Semi-supervised fuzzy clustering (SSFC) extends fuzzy clustering by incorporating semi-supervised learning. SSFC integrates the flexibility of fuzzy logic with the guidance of partial supervision [4–6]. Unlike hard clustering, which assigns each data point to exactly one cluster, fuzzy clustering allows each point to belong to multiple clusters with varying degrees of membership. This feature is especially suitable for problems with unclear cluster boundaries or noisy data.

Within the SSFC framework, the selection of high-quality seed instances plays a decisive role in clustering performance [7]. Well-chosen seeds can reduce labeling costs, enhance cluster separability, and stabilize convergence, whereas poorly chosen seeds may propagate errors, degrade clustering accuracy, and increase reliance on expert intervention. Despite its importance, seed selection remains a challenging and underexplored problem, requiring methods that are both theoretically sound and

 doi:10.20944/preprints202509.1405.v1

practically efficient. Thanks to its ability to effectively exploit a small amount of prior knowledge to guide the learning process, SSFC has attracted significant attention from the research community and has been widely applied in domains such as pattern recognition, information processing, and healthcare [2,6,8–13].

The difficulty of selecting good seeds arises from three main factors: (i) the non-trivial task of identifying informative patterns within uncertain cluster boundaries; (ii) the presence of noisy and imbalanced data, which can distort the clustering process; and (iii) the limited labeling budget, which necessitates efficient and adaptive query strategies.

Several studies have attempted to improve SSC through different strategies, including graph-based propagation, constraint expansion, and probabilistic seed selection [7,13]. While these methods have achieved certain successes, they still share common limitations: reliance on randomly chosen seeds or predefined constraints, lack of mechanisms to actively evaluate seed reliability, and insufficient robustness when dealing with noisy or overlapping clusters.

Fuzzy Min–Max Neural Networks (FMNN) have recently garnered increased attention in semi-supervised learning research, thanks to their ability to make effective use of limited labeled data. FMNN functions by dividing the data space into HXs. The visualization in Figure 1 illustrates the distribution of 2D HXs belonging to seven clusters on the Aggregation dataset [16], as partitioned by FMNN.

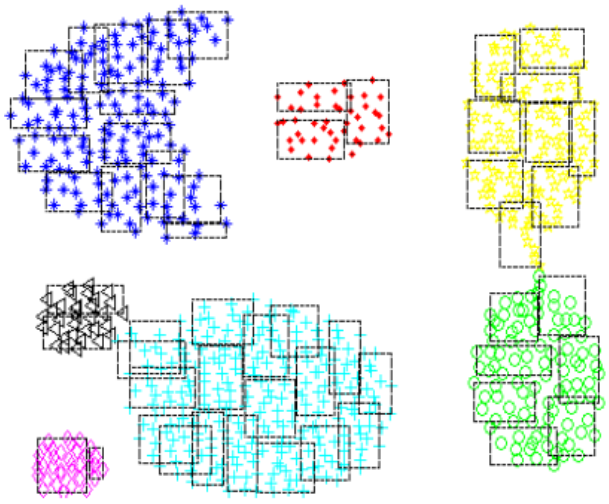


Figure 1. Illustration of fuzzy HXs in a two-dimensional Aggregation dataset, partitioned by the FMNN model.

Based on FMNN, the models such as GFMNN (General FMNN) [17], SS-FMM [2], SCFMN (SSC in FMNN) [11], MSCFMN (Modified SCFMN) [6], and WHFMN (Weighted Hyperboxes in FMNN) [18] have demonstrated the extensibility of FMNN in SSC tasks. However, approaches based on FMNN variants, although demonstrating flexibility in partitioning data through HXs, remain constrained by their dependence on seed quality and the absence of active querying strategies. These limitations highlight an important research gap: how to effectively select critical data points for labeling in order to reduce dependence on labeled data while maintaining clustering accuracy.

Although FMNN-based approaches have certain limitations, they remain a promising research direction with considerable potential and applicability [6]. In this study, we introduce a SSC model built upon FMNN and enhanced with an active seed selection strategy, referred to as FA-Seed. Rather than selecting seeds randomly, the model prioritizes data points that are expected to yield the most informative labels. This design aims to maximize information gain from each expert query while simultaneously minimizing the overall labeling effort.

The proposed model is applied to benchmark datasets to demonstrate its effectiveness in handling clustering problems characterized by fuzzy boundaries, class imbalance, and scarce labeled data. The main contributions of this work include:

 doi:10.20944/preprints202509.1405.v1

- Proposing an active seed selection strategy based on FMNN;
- Evaluating seed quality under limited label availability;
- Providing label suggestions to experts along with an optimized validation mechanism;
- Assessing performance on benchmark datasets.

The remainder of this paper is organized as follows: Section 2 presents the background and related work. Section 3 describes the proposed model in detail. Section 4 reports the experimental results. Finally, Section 5 discusses the conclusions and future research directions.

2. Related Works

In [7], the research team developed SSGC (Semi-Supervised Graph-based Clustering), an active learning algorithm that identifies seed instances using the k-nearest neighbor and min–max algorithms. Its key idea is to construct a weighted k-NN graph based on shared neighbors and then partition it into connected components under a cut condition. Although its label propagation mechanism—combining an adaptive cut threshold (θ) with graph connectivity—is efficient and flexible for varying cluster shapes, seed selection in SSGC remains random, making performance sensitive to initial labels.

Beyond graph-based approaches, Xin Sun et al. [13] proposed a constraint-driven framework consisting of CS-PDS (Constraints Self-learning from Partial Discriminant Spaces) and FC2 (Finding Clusters on Constraints). CS-PDS iteratively expands initial constraints using an Expectation–Maximization cycle, while FC2 performs clustering on the enriched constraint set. This approach achieves strong performance on small-scale data but relies heavily on the representativeness of initial constraints and lacks an explicit active query mechanism.

A probabilistic alternative was introduced by Bajpai et al. [12], who employed Determinantal Point Processes (DPP) to optimize seed diversity. Unlike SSGC, DPP selects seeds that are both representative and broadly distributed, which are then used as K-Means centroids, reducing convergence to poor local optima. However, this approach incurs high computational cost ($O(n^3)$) and struggles with noisy or overlapping clusters due to the absence of propagation or denoising mechanisms.

Another important line of research builds on FMNN, which leverages HXs to represent uncertainty and handle limited supervision. Among these approaches, MSCFMN stands out by refining HX construction and overlap handling. In MSCFMN, the number of HXs is fixed to the number of clusters; they expand under a threshold θ, with contractions applied to resolve overlaps. While such refinements improve accuracy over earlier FMNN-based models, they still suffer from rigid partitioning of the feature space, the absence of active seed selection, and high sensitivity to the threshold parameter θ.

Our proposed FA-Seed extends MSCFMN by retaining its HX-based partitioning mechanism while introducing adaptive seed selection, seed-quality scoring, and few-shot generalization. These enhancements address the limitations of MSCFMN and enable FA-Seed to achieve superior clustering performance under limited supervision.

3. Proposed Method: FA-Seed

3.1. The Idea of Proposed Model

The effectiveness of SSC depends strongly on the quality of selected seeds. Poor seeds can propagate errors, while reliable seeds reduce labeling costs and improve clustering accuracy. However, selecting high-quality seeds is challenging in practice due to noisy data, overlapping clusters, and limited supervision. To address this, the proposed FA-Seed model extends the MSCFMN framework [6] by integrating HX partitioning, active querying, and few-shot generalization.

Built on the MSCFMN, FA-Seed represents clusters as HXs, naturally handling uncertainty and overlapping regions. Unlike random or fixed initialization, the model adaptively scores candidate seeds using density, centrality, and uncertainty, and selectively queries the most informative samples. Verified seeds then guide controlled label propagation through the HX structure, ensuring consistency and reducing error propagation.

 doi:10.20944/preprints202509.1405.v1

Unlike MSCFMN, which automatically assigns a single labeled point to each HX based on data distribution characteristics, FA-Seed incorporates an active learning mechanism to purposefully evaluate and select potential seed candidates. Rather than relying on default labeling, the model actively queries and verifies informative data points with experts. Additionally, FA-Seed integrates various criteria such as centroid deviation, data density, and uncertainty to filter and assess seed quality, thereby enhancing the model's generalization ability and clustering accuracy under limited labeled data conditions.

Figure 2(a) visually illustrates the limitation of the MSCFMN model during Phase 1 — the phase of defining auxiliary information (seed selection) — on the Aggregation dataset. The Aggregation dataset consists of 788 samples grouped into 7 clusters [16], as shown in Figure 2(a).

In MSCFMN, the number of HXs generated during this initial phase is constrained to match the number of clusters. Specifically, only 7 HXs are created, as seen in Figure 2(b). However, several of these HXs (e.g., HXs 1, 3, 4, and 5) overlap with multiple clusters, reducing the reliability of label assignment for the corresponding seed points. If seed points from such ambiguous HXs are selected and used in label propagation, the model may suffer from reduced accuracy due to mislabeling. Furthermore, this restricted partitioning approach limits the model's ability to capture more complex or uneven data distributions within individual clusters. MSCFMN is thus highly influenced by both the number and shapes of clusters.

In contrast to MSCFMN, FA-Seed addresses these limitations by enabling a more flexible partitioning of the input space, allowing the generation of a greater number of HXs beyond the initial number of clusters. This adaptive strategy improves coverage and granularity. As illustrated in Figures 2(c)–(d):

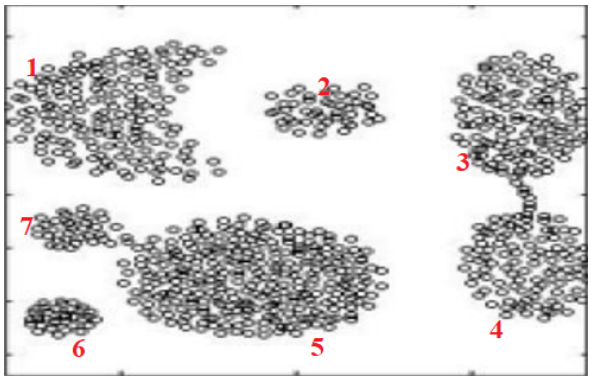
- Figure 2(c): FA-Seed generates 17 HXs and selects 13 data points for expert query, accounting for 1.65% of the total samples.
- Figure 2(d): 26 HXs with 21 queries (2.66%).
- Figure 2(e): 38 HXs with 29 queries (3.68%).
- Figure 2(f): 52 HXs with 40 queries (5.08%).

Beyond simply expanding the number of HXs, FA-Seed incorporates a quality assessment mechanism to enhance label query efficiency. Specifically, only HXs with balanced and representative data distributions are selected for expert queries. HXs with skewed distributions or very few data points are excluded to avoid wasting labeling budget and to reduce the risk of incorrect label propagation.

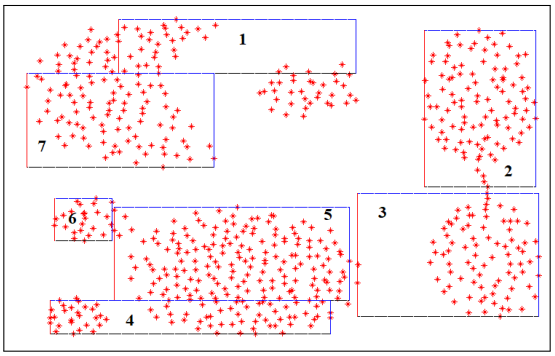
Figure 3 illustrates the process of calculating and selecting HXs in the proposed FA-Seed model. HXs with a high level of data balance are selected (labeled numerically from 1 to 21), whereas imbalanced HXs are excluded from selection (denoted by characters a, b, c, d, and e).

Thanks to these two improvements — adaptive HX generation and pre-query quality evaluation — FA-Seed enables more effective learning in SSC tasks involving unlabeled and non-uniform data. It also enhances clustering accuracy while reducing the model's dependence on prior assumptions about the number or shape of clusters.

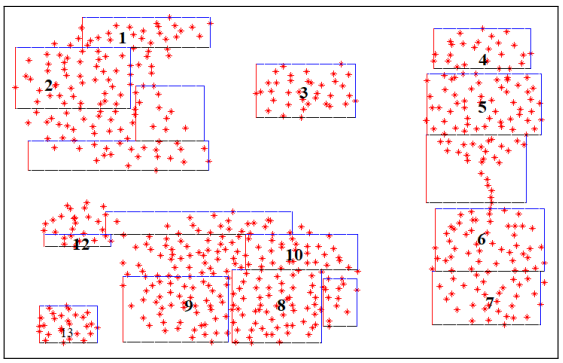
 doi:10.20944/preprints202509.1405.v1



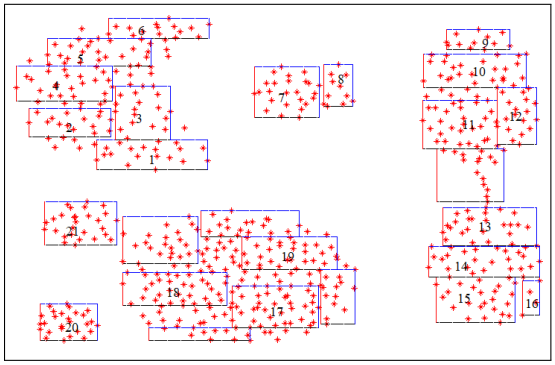
(**a**)



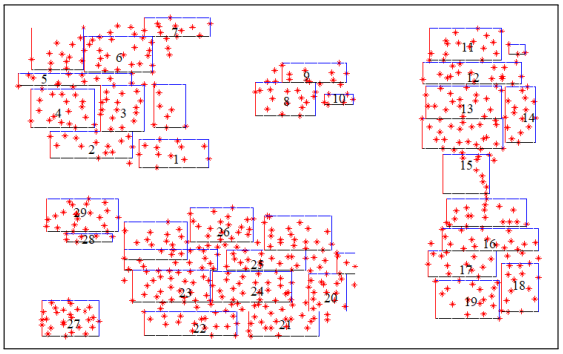
(**b**)



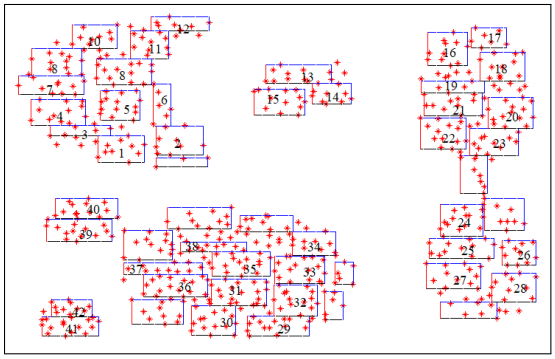
(**c**)



(**d**)



(**e**)



(**f**)

Figure 2. Illustration of HX partitioning on the Aggregation dataset by FMNN-based models, including: (**a**) Data distribution of the Aggregation dataset. (**b**) Data space partitioned into 7 HXs by MSCFMN (one per cluster). (**c**) Data space partitioned into 17 HXs by FA-Seed with 13 queried points. (**d**) Data space partitioned into 26 HXs bay FA-Seed with 21 queried points. (**e**) Data space partitioned into 38 HXs by FA-Seed with 29 queried points. (**f**) Data space partitioned into 52 HXs by FA-Seed with 40 queried points.

 doi:10.20944/preprints202509.1405.v1

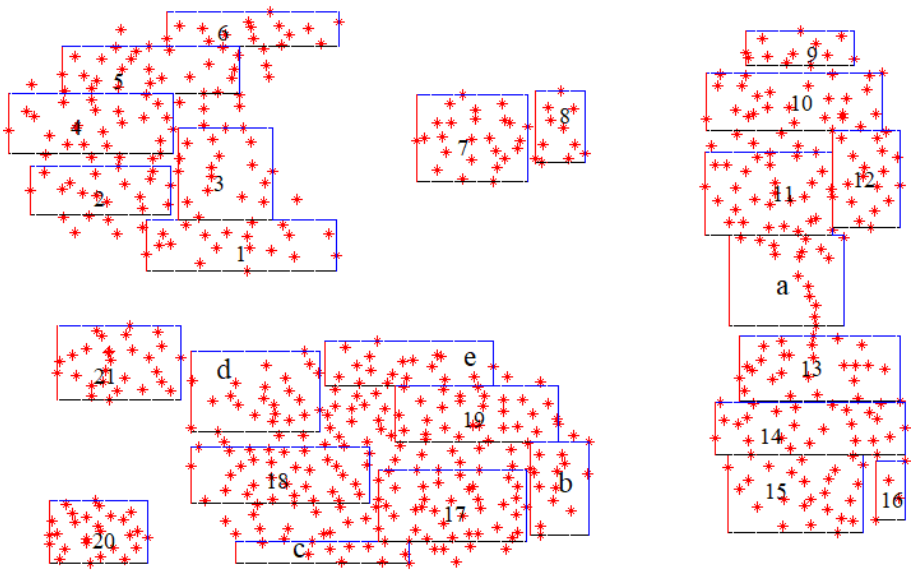


Figure 3. HX selection on the **Aggregation** dataset using FA-Seed. Balanced HXs (1–21) are retained, whereas imbalanced HXs (a–e) are discarded.

3.2. Setup and Definitions

Fuzzy hyperbox and membership. Assume we have M data instances $X = \{x_i\}_{i=1}^{M} \subset \mathbb{R}^n$, where each $x_i \in \mathbb{R}^n$, which can be mapped into G ground-truth clusters. Let the index set $\{1, \ldots, M\}$ be partitioned into a labeled subset $\mathcal{L}$ and an unlabeled subset $\mathcal{U}$, with $\mathcal{L} \cup \mathcal{U} = \{1, \ldots, M\}$ and $\mathcal{L} \cap \mathcal{U} = \varnothing$, where the split is induced by partitioning the input space into HXs.

Let $B = \{B_j\}_{j=1}^{K}$ denote the current set of HXs (so $K = |B|$). Denote the labeled and unlabeled subsets by B_{lab} and B_{unl}, respectively, with

$$B = B_{\text{lab}} \uplus B_{\text{unl}} \iff B_{\text{lab}} \cup B_{\text{unl}} = B, \;\; B_{\text{lab}} \cap B_{\text{unl}} = \varnothing. \tag{1}$$

A fuzzy HX $B_j = (V_j, W_j)$ is as defined in (2).

$$B_j = (V_j, W_j), \quad V_j, W_j \in \mathbb{R}^n, \; V_j \leq W_j \tag{2}$$

where V_j and W_j are the minimum and maximum vertices of the HX in the input space.

Each HX B_j is associated with a class label $c_j \in \{1, \ldots, C\}$, which is assigned during the learning process. For $i \in \mathcal{L}$, labels $y_i \in \{1, \ldots, C\}$ are known; for $i \in \mathcal{U}$, predicted labels are denoted by $\hat{y}_i$.

Let x be a candidate sample. Its membership to B_j is $\mu_j(x, B_j) \in [0, 1]$, as defined in Equation (3). Given the current set of HXs $B = \{B_j\}_{j=1}^{K}$, the HX with the largest membership value for x is determined by Equation(5).

$$\mu_j(x, B_j) = \frac{1}{n} \sum_{r=1}^{n} \left[1 - f(x_r - w_{jr}, \gamma) - f(v_{jr} - x_r, \gamma)\right] \tag{3}$$

where the function $f(.,.)$ is a two-parameter function used to modulate the attenuation of the membership value μ_j, based on the correlation between the data sample x and the boundary of the HX B_j. The sensitivity parameter γ directly affects the shape of this attenuation: a larger γ leads to faster degradation of the membership value, thereby making the model more responsive to boundary violations. Its mathematical formulation is given in Equation (4):

$$f(x, y) = \begin{cases} 0, & \text{if } x \times y < 0 \\ x \times y, & \text{if } 0 \leq x \times y \leq 1 \\ 1, & \text{otherwise} \end{cases} \tag{4}$$

 doi:10.20944/preprints202509.1405.v1

$$j^\star \;=\; \arg \max_{j\in\{1,\dots,K\}} \mu_j(x, B_j). \tag{5}$$

Seed set based on centers and purity. Let T_j be the core set used to compute the data center d_j as in (6) (the τ-core $T_j = \{\, i : \mu_j(x_i, B_j) \geq \tau \,\}$). The geometric center g_j is defined in (7), and the center deviation e_j is defined in (8). A HX is regarded as pure if e_j satisfies constraint (9) is satisfied.

$$d_j \;=\; \frac{1}{|T_j|} \sum_{i\in T_j} x_i. \tag{6}$$

$$g_j \;=\; \tfrac{1}{2}\,(V_j + W_j). \tag{7}$$

$$e_j \;=\; \|\, g_j - d_j \,\|_2. \tag{8}$$

$$e_j \;\leq\; \delta, \qquad \delta > 0 \text{ is user-defined.} \tag{9}$$

Consequently, J_Q is the index set of pure fuzzy HXs, given by (10); the candidate query set Q is the collection of their data centers, as in (11); and S denotes the set of fully contained samples, defined in (12) and generated from Q.

$$J_Q = \{\, j \mid B_j \in B, e_j \leq \delta \,\} \tag{10}$$

$$Q \;=\; \{\, d_j \mid\; j \in J_Q \,\}. \tag{11}$$

$$S \;=\; \bigcup_{j\in J_Q} \{\, x_i \in X \,\big|\, \mu_j(x_i, B_j) \geq \tau \,\}, \qquad \delta > 0,\ \tau \in [0,1]. \tag{12}$$

Adaptive partition and assignment with fuzzy HXs. Given the current HXs $B = \{B_j\}_{j=1}^{K}$ and a candidate sample x, first select the HX with the largest membership (as defined in Equation (5)). Update the space of HXs in the set B according to:

- *Containment rule:* If $\exists\, j$ such that $\mu_j(x, B_j) = 1$, assign x to that B_j (uniqueness follows from $\Omega(B_a, B_b) \leq 0$).
- *Expansion:* Form the tentative expansion of $B_{j^\star}$ as defined in Equation (13). Accept it and set $B_{j^\star} \leftarrow B'_{j^\star}$ *iff* all of the following hold:

  (1) size constraint $\theta(x, B_{j^\star}) \leq \theta_{\max}$ (as defined in Equation (14));
  (2) after any necessary contraction, the overlap constraint $\Omega(B_i, B_{j^\star})$ satisfies the constraint in Equation (16).

  If accepted, assign x to the updated $B_{j^\star}$.
- *Create a new HX:* If the expansion is not accepted, initialize a new HX around x as in Equation (17).

$$B'_{j^\star} = \big(\min(V_{j^\star}, x),\ \max(W_{j^\star}, x)\big). \tag{13}$$

$$\theta(x, B_j) \leq \theta_{\max}, \qquad \theta_{\max} \text{ is user-defined.} \tag{14}$$

where $\theta(x, B_j)$ denotes the post-expansion size of B_j (e.g., as in (15)):

$$\theta(x, B_j) = \frac{1}{n} \sum_{r=1}^{n} \sqrt{\big(|\max(x_r, W_{jr})| - |\min(x_r, V_{jr})|\big)^2}. \tag{15}$$

 doi:10.20944/preprints202509.1405.v1

$$\Omega(B_i, B'_j) \leq 0, \qquad \forall\, i \neq j, \tag{16}$$

where $\Omega(\cdot,\cdot)$ is the overlap measure.

$$B_{\text{new}} = (x, x). \tag{17}$$

Fuzzy HX-level labeling by inheritance. With a decreasing schedule $\{\beta_t\}_{t=0}^{T}$ ($\beta_0 > \cdots > \beta_T \geq \beta_{\min}$), an unlabeled HX B_j inherits the label of the nearest *labeled* HX iff both gates hold: (i) the center-alignment gate $e_j \leq \delta$ (with e_j defined in Equation (8)); and (ii) the membership gate

$$\mu_{j^\star}(d_j, B_{j^\star}) \;\geq\; \beta_t, \qquad \beta_t \in [0,1]. \tag{18}$$

where d_j is the data center defined in Equation (6) and $j^\star$ is chosen by Equation (19); labels are then updated according to Equation (20).

$$j^\star \;=\; \arg\max_{\ell:\, B_\ell \in B_{\text{lab}}} \mu_\ell(d_j, B_\ell). \tag{19}$$

When both gates are satisfied, update the label and the sets:

$$\big(e_j \leq \delta\big) \;\wedge\; \big(\mu_{j^\star}(d_j, B_{j^\star}) \geq \beta_t\big) \;\Rightarrow\; \begin{cases} \text{label}(B_j) \leftarrow \text{label}(B_{j^\star}), \\ B_{\text{lab}} \leftarrow B_{\text{lab}} \cup \{B_j\}, \\ B_{\text{unl}} \leftarrow B_{\text{unl}} \setminus \{B_j\}. \end{cases} \tag{20}$$

Iterate (20) for $t = 0, \ldots, T$. Finally, submit the remaining unlabeled (outlier/atypical) HXs for expert labeling:

$$Q \;=\; B_{\text{unl}}. \tag{21}$$

Edge cases: if $B_{\text{lab}} = \varnothing$, skip or defer inheritance or query an expert. If no HX exists yet ($K = 0$), initialize B_1 defined as in Equation (17); if the current sample x has a ground-truth label, set $c_1 = y_i$, otherwise leave it pending.

Training procedure:

Stage 1 (one-shot): query all seeds $Q^{(1)} = S$ (with S defined in Equation (11)) to obtain $(B_{\text{lab},0}, B_{\text{unl},0})$.

Stage 2 (progressive): with a decreasing schedule $\{\beta_t\}$, for each $u \in B_{\text{unl},t}$, let $j^\star(u)$ is selected according to Equation (19).

If the alignment gate $e_u \leq \delta$ ((9)) and the membership gate $\mu_{j^\star(u)}(d_u, B_{j^\star(u)}) \geq \beta_t$ ((18)) both hold, then move u to $B_{\text{lab},t+1}$; otherwise keep it in $B_{\text{unl},t+1}$. Finally, query the remainder

$$Q^{(2)} = B_{\text{unl},T+1}. \tag{22}$$

3.3. Model Architecture and Learning Process

3.3.1. Architecture of FA-Seed Model

The FA-Seed model combines the architecture of the FMNN [19,20] with a semi-supervised clustering strategy, organized into two main phases for clustering on the dataset X. In the first phase, FA-Seed identifies auxiliary information (labels) for a subset of the data, denoted as S, while the remaining unlabeled samples form the subset U ($X = S \cup U$). In the second phase, FA-Seed propagates labels from S to U by constructing and adapting HXs, thereby forming clusters for the entire dataset X.

As illustrated in Figure 4, the architecture of FA-Seed consists of two main components:

- Seed scoring module: Definition of auxiliary information based on the partitions the data space into HXs, evaluates candidate seeds using criteria such as data density, centroid deviation, and uncertainty, and assigns labels to reliable seeds within their respective HXs.

 doi:10.20944/preprints202509.1405.v1

- Controlled label propagation and extended evaluation: propagates labels from the verified seed set to the unlabeled data, ensuring consistency and accuracy across clusters. This component leverages the HX structure to minimize error propagation and actively queries uncertain samples that cannot be confidently assigned to a cluster.

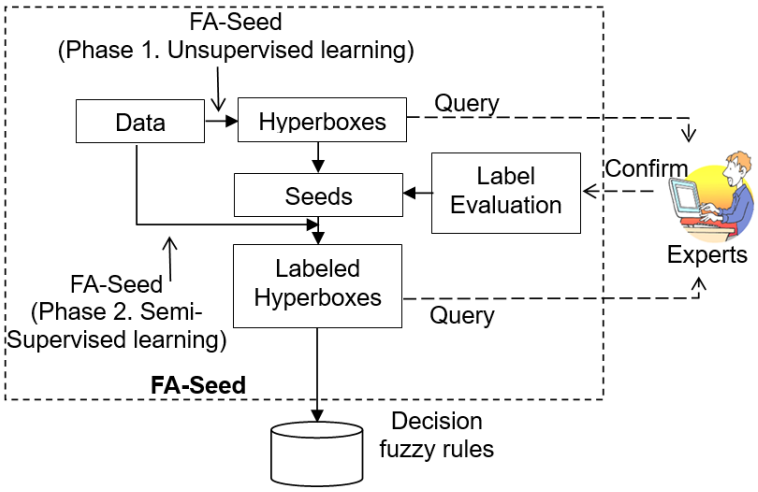


Figure 4. The architecture of the proposed FA-Seed model with an integrated label evaluation component.

3.3.2. Neural Network Architecture of FA-Seed

The FA-Seed model is structured as a four-layer neural network, as illustrated in Figure 5. We consider four layers $\mathsf{F}_A \to \mathsf{F}_B \to \mathsf{F}_H \to \mathsf{F}_C$, each playing a distinct role in the semi-supervised clustering process:

- **Input layer** F_A**.** This layer contains n nodes, each representing an input attribute. A sample is $x_i = (x_{i1}, x_{i2}, \ldots, x_{in})$ and is mapped into HXs in F_B using the min–max boundary vectors v and w.
- **HX scoring layer** F_B**.** Includes K nodes (HXs) produced at Stage 1, each corresponding to a HX $B_j = (v_j, w_j) \in B$. These HXs are constructed incrementally and capture the local data structure in the input space. The activation is the membership

$$b_j(x_i) = \mu_j(x_i, B_j) \in [0, 1]. \tag{23}$$

- **Evaluation/propagation layer** F_H**.** HX nodes are partitioned into labeled and unlabeled sets: B_{lab} (size K_{lab}) and B_{unl} (size K_{unl}). This layer applies the gates and the inheritance rule to move HXs from B_{unl} to B_{lab} (at termination, $B_{\text{unl}} = \varnothing$ so $K_{\text{unl}} = 0$ and typically $K_{\text{lab}} > K$).
- **Output layer** F_C: Contains G nodes, each representing one output cluster. The binary relationship between HX $L_j \in F_H$ and class $c_g \in F_C$ is represented by the matrix $U = \{u_{jg}\}$:

$$u_{jg} = \begin{cases} 1, & \text{if } l_j \in c_g, \\ 0, & \text{otherwise.} \end{cases} \tag{24}$$

The membership score of a sample to class c_g is computed as:

$$c_g = \max_{1 \le j \le q} l_j \cdot u_{jg}, \tag{25}$$

where l_j is the activation of HX j. The set of HXs defining class g is given by:

$$C_g = \bigcup_{j \in G} L_j, \tag{26}$$

with G denoting the index set of HXs assigned to class g.

 doi:10.20944/preprints202509.1405.v1

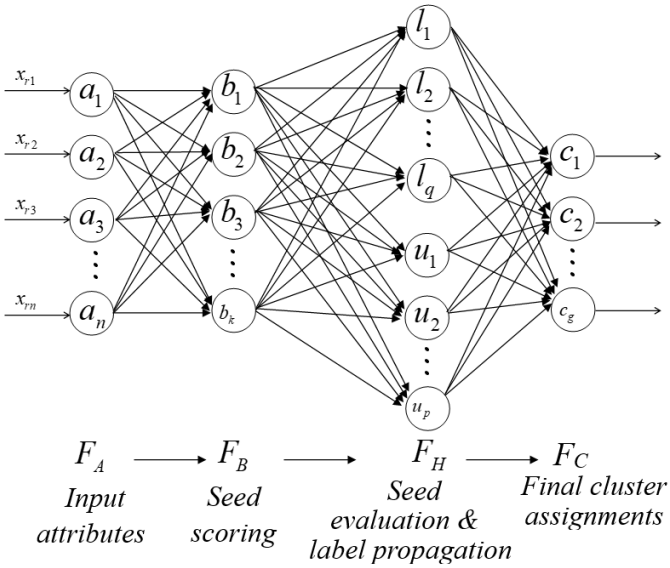


Figure 5. Neural Network Architecture of FA-Seed.

3.3.3. Learning Algorithm

The first stage of the learning algorithm defines auxiliary information based on the partition of the data space into fuzzy HXs. Candidate seeds are evaluated using criteria such as data density, centroid deviation, and uncertainty, and reliable seeds are then assigned labels within their respective HXs. The learning algorithm for this phase is presented in Algorithm 1.

Algorithm 1: Seed Set Construction Phase of FA-Seed

Input: Dataset X ($M = |X|$); $\theta_{\max}$; deviation threshold δ

Output: Updated data set X, B

Step 1. Hyperbox generation

Perform one pass over X to construct fuzzy HX set B ($B = \{B_1, \ldots, B_j\}$, ($j \in K$, $K = |B|$)) using min–max boundaries with θ_{max}.

Step 2. Hyperbox filtering (Identify data instances to query)

For each HX $B_j \in B$, denote by $x_i \subseteq X$ the set of samples assigned to B_j, and by $V_j, W_j \in \mathbb{R}^d$ its min/max vertices.

Compute the *data center* defined by Equation (6) and the *geometric center* of the HX B_j defined by Equation (7).

Keep only those B_j whose centroid is sufficiently close to its geometric center, and remove B_j if condition (8) is violated: $B := B \setminus \{B_j\}$.

Step 3. Seed set construction, partition and reordering

Build the seed set S from samples fully contained in at least one surviving HX by by Equation (12).

Build a permutation π of $\{1, \ldots, m\}$ that places S first and D after (X is the set of unlabeled samples, i.e., points that do not fully belong to any HX), and form the reordered dataset $X' = (X_{\pi(1)}, \ldots, X_{\pi(m)})$.

return *B (the surviving HXs), X' (the reordered dataset)*

* *Computational Complexity of FA-Seed – Algorithm 1*:

$$T_1 = O(nMK). \tag{27}$$

The second stage of the learning algorithm describes the propagation of labels from the verified seed set to the unlabeled data, ensuring consistency and accuracy across clusters. By leveraging the HX

 doi:10.20944/preprints202509.1405.v1

structure, this stage minimizes error propagation and actively queries uncertain samples that cannot be confidently assigned to a cluster. The learning algorithm for this stage is presented in Algorithm 2.

$$\mu_j(x, L_j) = \max\{\mu_j(x, L_j) : j = 1, \ldots, k\}, \tag{28}$$

where $\theta(x, L_j)$ denotes the expansion measure when adding x to L_j.

$$\max_{j \in \{1,\ldots,|L|\}} \mu(H_{\text{new}}, L_j) \geq \beta, \tag{29}$$

$$U_l^{\text{label}} = L_{j^\star}^{\text{label}}, \quad j^\star = \arg \max_{j \in \{1,\ldots,|L|\}} \mu(g_l, L_j), \tag{30}$$

$$\begin{cases} L := L \cup \{U_j\} \\ U := U \setminus \{U_j\} \end{cases} \tag{31}$$

* *Computational Complexity of FA-Seed – Algorithm 2*:

$$T_2 = O(nMK). \tag{32}$$

Computational complexity of the FA-Seed algorithm:

$$T = T_1 + T_2 = O(nMK). \tag{33}$$

Algorithm 2: Guided Hyperbox Expansion with Label Propagation

Input: Reordered dataset X; S; β; δ; $\theta_{\max}$.
Output: L.
for $x \in X$ **do**
  Compute $\{\mu_j(x, L_j)\}_{L_j \in \mathcal{L}}$ and set $j^\star \leftarrow \arg\max_j \mu_j(x, L_j)$;
  if $\theta(x, L_{j^\star}) \leq \theta_{\max}$ **then**
    Expand $L_{j^\star}$ with x (update $V_{j^\star}, W_{j^\star}$, resolve overlaps if needed);
    // $\theta(x, L_{j^\star})$ calculated according to Equation (15)
  else
    Create a new HX H_{new} initialized by x;
    if $x \in S$ **then**
      label$(H_{\text{new}}) \leftarrow$ label(x);
      Insert H_{new} into L ($L := L \cup \{H_{\text{new}}\}$)
    else
      Find $L_j \in L$ by Equation (28) such that the constraint in Equation (29) holds;
      if *such an* L_j *exists* **then**
        label$(H_{\text{new}}) \leftarrow$ label(L_j);
        Insert H_{new} into L ($L := L \cup \{H_{\text{new}}\}$)
      else
        Insert H_{new} into the unlabeled set U ($U := U \cup \{H_{\text{new}}\}$)
for *each* $U_\ell \in U$ **do**
  Compute the centroid g_ℓ of U_ℓ as defined in Equation (6);
  Find $L_j \in L$ such that $\mu(g_\ell, L_j)$ satisfies Equation (28) and $\max_{j \in \{1,\ldots,|L|\}} \mu(g_\ell, L_j) \geq \beta$;
  if U_ℓ *satisfies the conditions* **then**
    label$(U_\ell) \leftarrow$ label(L_j);
    Move U_ℓ from U to L calculated according to Equation (31).
  else
    Suggest label(U_ℓ) using Equation (30) based on g_ℓ;
return L

 doi:10.20944/preprints202509.1405.v1

4. Experiments

4.1. Experimental Objectives

The goal of this section is to evaluate the effectiveness of the proposed FA-Seed model in semi-supervised clustering tasks. Specifically, we aim to investigate: (i) the ability of FA-Seed to improve clustering accuracy under limited labeled data, (ii) its efficiency in seed selection compared to baseline methods, and (iii) its applicability to benchmark datasets [16,21].

To evaluate the performance of FA-Seed, experiments were conducted to compare it against several state-of-the-art semi-supervised clustering baselines, including SSDBSCAN, SSK-Means, K-Means, SSGC [7], MSCFMN [6], and FC2 [13].

4.2. Experimental Setup

4.2.1. Datasets

We conduct experiments on benchmark datasets to evaluate the general performance and applicability of the proposed model. The benchmark datasets are drawn from standard UCI and CS collections, including Iris, Wine, Zoo, Soybean, Ecoli, Yeast, PID, Spiral, Pathbased, Thyroid, and others. These datasets span a wide range of sample sizes, dimensions, and cluster structures, thereby providing a comprehensive basis for evaluating clustering quality under diverse conditions. The detailed characteristics are summarized in Table 1. All datasets are normalized prior to experimentation. For datasets with missing values, imputation is performed following the approach of Batista [22], where missing entries are replaced with the mean value of the corresponding attribute, as defined in Equation (34).

$$A_{hj} = \frac{1}{m} \sum_{i}^{m} A_{ij}, \tag{34}$$

Table 1. Benchmark datasets used in the experiments.

Dataset	#Instances	#Features	#Clusters
Soybean	47	35	4
Zoo	101	16	7
Iris	150	4	3
Wine	178	13	3
Thyroid	215	5	3
Flame	240	2	2
Spiral	312	2	3
Pathbased	317	2	3
Ecoli	336	8	8
Jain	373	2	2
PID	768	8	2
Aggregation	788	2	7
Yeast	1484	8	10
ThyroidDisease	7200	21	3

4.2.2. Parameter Settings

The following settings are used unless otherwise specified: fuzziness parameter $\gamma = 10$, initial threshold $\beta = 0.99$, threshold decay factor $\phi = 0.9$, and deviation threshold θ_δ selected empirically based on the error correction method.

4.2.3. Evaluation Metrics

As noted earlier, clustering performance depends on both the quality and the number of selected seeds. We therefore conduct comparative experiments on benchmark datasets, comparing against state-of-the-art semi-supervised clustering baselines and evaluating performance using Rand Index

 doi:10.20944/preprints202509.1405.v1

(RI), accuracy, training time, number of queries (NoQ), and number of seeds (NoS) obtained via expert querying. RI is computed by (35), and accuracy by (36).

$$RI_{(X,Y)} = \frac{2(u+v)}{n(n-1)} \tag{35}$$

$$Accuracy = \frac{1}{n}\sum_{i=1}^{n} H(y_i = \overline{y_i}) \tag{36}$$

where:

- u is the number of correct decisions when $y_i = y_j$ in both clustering and ground truth.
- v is the number of correct decisions when $y_i \neq y_j$ in both clustering and ground truth.
- n is the total number of samples.
- $H(y) = 1$ if the prediction is correct; otherwise $H(y) = 0$.

4.3. Results

As analyzed, increasing the proportion of seeds often contributes to improving clustering quality. In the first experiment, we focus on evaluating the ability to obtain seeds through expert queries.

Figure 6 visually illustrates the comparison of the number of seeds obtained between the methods when using the same number of queries. The results show that, at the same NoQ level, FA-Seed consistently obtains a higher number of seeds compared to MSCFMN [6]. Notably, FA-Seed achieves a high seed ratio even when the number of queries is low. This is a significant advantage, as it greatly reduces manual labeling costs. In addition, a high seed ratio directly contributes to enhancing the model's performance. The main reason FA-Seed achieves a larger number of seeds is its finer HX partitioning mechanism compared to MSCFMN, which allows the model to approach cluster boundaries more closely and thus more effectively exploit potential data points for labeling.

Figure 7 presents the comparison results of FA-Seed with SSGC, SSDBSCAN, SSK-Means, K-Means, MSCFMN [6,7], using the Rand Index (RI) to evaluate clustering quality. The seeds are randomly selected in each run, and the process is repeated 20 times, with the average RI value recorded for comparison. From the visual results, it can be observed that FA-Seed generally performs better or at least comparable to the other methods on most datasets. For datasets with a clear cluster structure such as Soybean, Zoo, and Iris, FA-Seed achieves near-perfect RI values and greater stability compared to other methods. This demonstrates its ability to effectively leverage seed information to assign labels accurately from the very beginning. For datasets with a large number of clusters and evenly distributed data such as Ecoli, FA-Seed still maintains high and stable RI values when the number of seeds changes. For datasets with a strong imbalance in cluster sizes such as Yeast (ranging from 5 to over 400 objects per cluster) or with high overlap levels such as Thyroid, FA-Seed achieves better results thanks to its optimal seed selection mechanism from "trusted HXs" instead of random selection, thereby reducing the negative impact of small or overlapping clusters.

 doi:10.20944/preprints202509.1405.v1

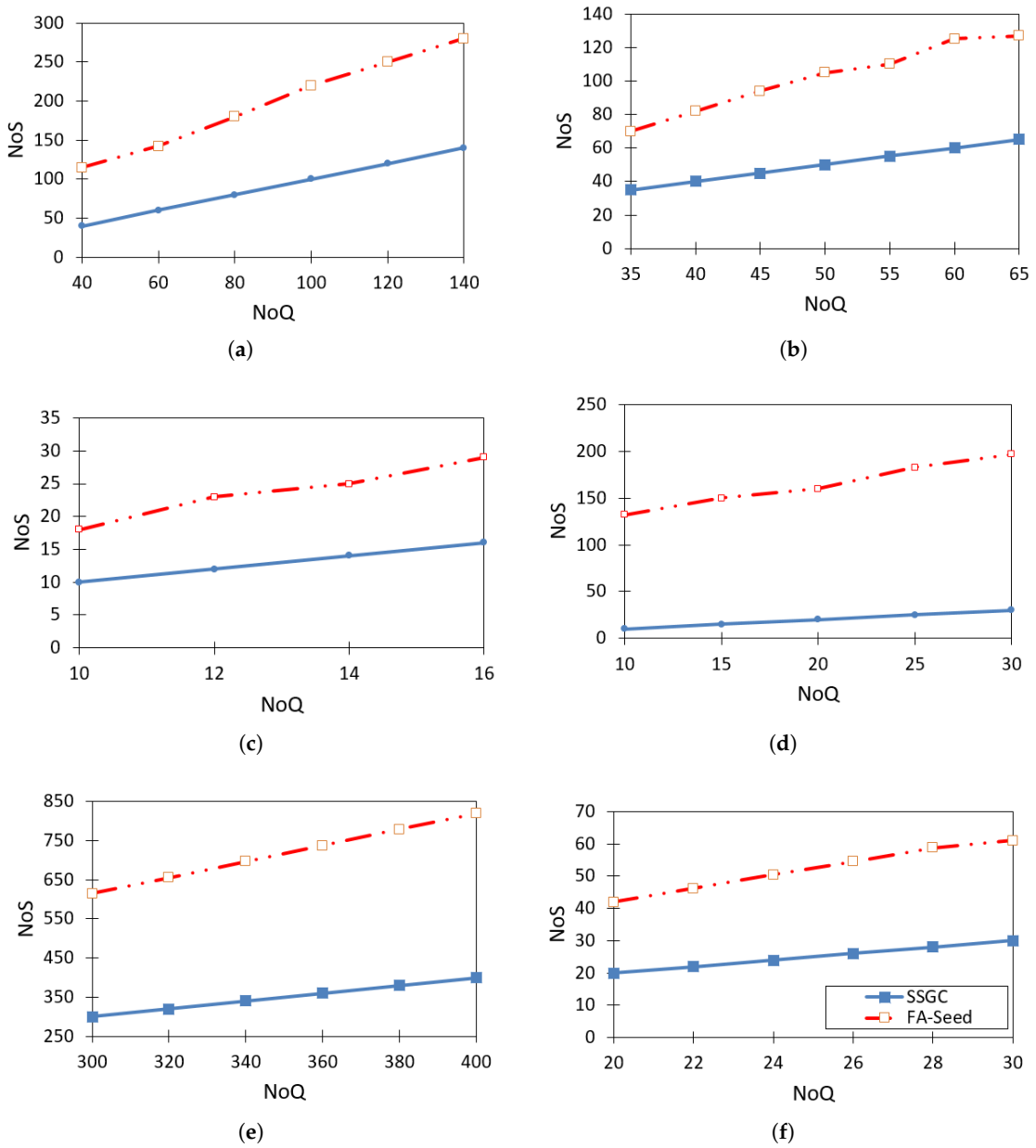


Figure 6. Comparison of NoS values between FA-Seed and baseline algorithms on benchmark datasets, including: (**a**) Ecoli dataset, (**b**) Iris dataset, (**c**) Soybean dataset, (**d**) Thyroid dataset, (**e**) Yeast dataset, (**f**) Zoo dataset.

 doi:10.20944/preprints202509.1405.v1

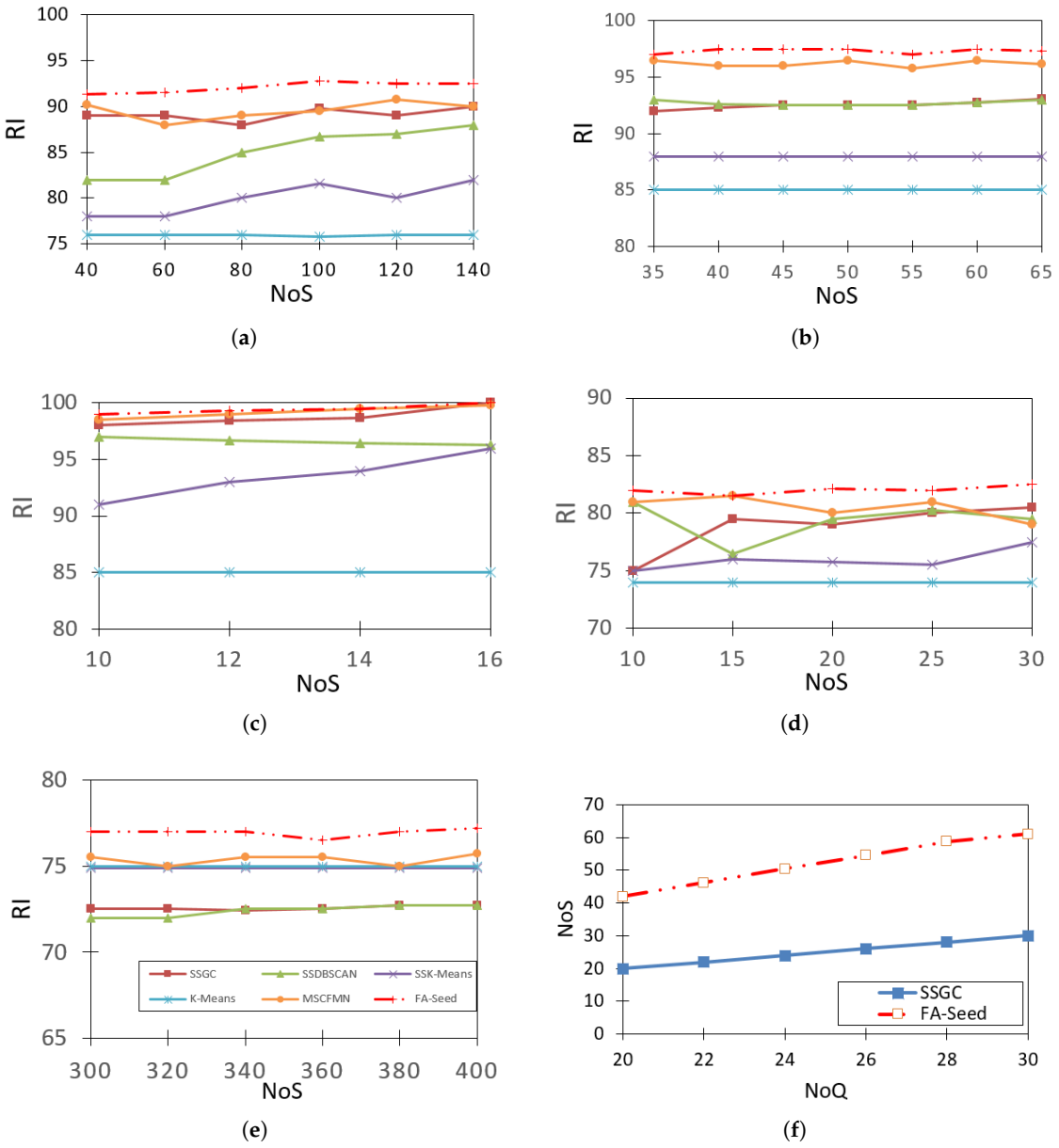


Figure 7. Comparison of RI values between FA-Seed and baseline algorithms on Benchmark datasets, including: (**a**) Ecoli dataset, (**b**) Iris dataset, (**c**) Soybean dataset, (**d**) Thyroid dataset, (**e**) Yeast dataset, (**f**) Zoo dataset.

Figure 8 is a graphic comparing the Accuracy measure between FA-Seed and MSCFMN. The results indicate that FA-Seed achieves accuracy comparable to or higher than MSCFMN on most datasets, with notable improvements on datasets with imbalanced class distributions (e.g., *Thyroid*, *Pathbased*, *Zoo*). This demonstrates that the high-quality seed selection mechanism in FA-Seed can enhance clustering performance even under challenging data conditions.

 doi:10.20944/preprints202509.1405.v1

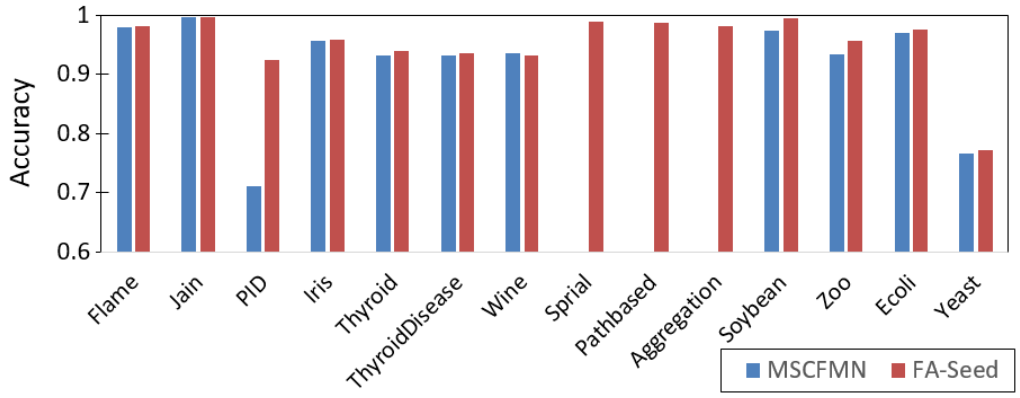
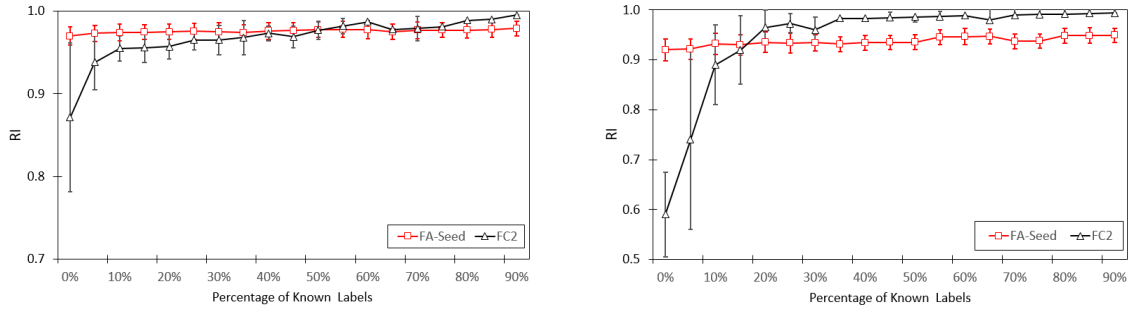


Figure 8. Visual comparison chart of Accuracy measures of FA-Seed and MSCFMN.

Figure 9 is a chart comparing the RI index between FA-Seed and FC2 [13]. The results show that FA-Seed consistently achieves higher RI values across all proportions of labeled data. Notably, the performance gap in RI between FA-Seed and FC2 is largest at lower labeling rates, confirming the advantage of the seed selection strategy in semi-supervised settings with limited labeled data.



(**a**) (**b**)

Figure 9. Average clustering performance RI achieved by the proposed FA-Seed and FC2, including: (**a**) Iris dataset, (**b**) Wine dataset.

Figure 10 presents a runtime comparison between FA-Seed and the baseline algorithms on the benchmark datasets. As shown, FA-Seed is consistently the fastest (or among the fastest). In terms of time complexity, FA-Seed has a computational cost of $O(nMK)$ (Equation (33)), where $M = |X|$ denotes the number of data samples, n is the number of attributes (data dimensions), and $K = |B|$ is the number of HXs. In practice, K is often significantly smaller than M, allowing FA-Seed to scale efficiently even as the dataset size grows.

In contrast, SSGC—the main algorithm presented in [7]—has a complexity of $\mathcal{O}(n^2)$ due to the need to construct a k-nearest neighbor graph and perform label propagation, which becomes costly for large datasets. SSDBSCAN achieves an average complexity of $\mathcal{O}(n \log n)$ when using optimal data structures, but still reaches $\mathcal{O}(n^2)$ in the worst case. For SSK-Means and semi-supervised K-Means, the complexity is $\mathcal{O}(t \times n \times c \times m)$, where t is the number of iterations and c is the number of clusters; the computational cost increases rapidly when either t or c is large.

Overall, FA-Seed demonstrates a clear advantage by partitioning the dataset into HXs and identifying "reliable HXs" to select high-quality seeds, while obtaining a larger number of seeds in a single independent stage. This approach allows users to proactively determine the number of labels based on the number of reliable HXs generated, thereby significantly reducing the labeling effort while maintaining, or even improving, clustering quality.

 doi:10.20944/preprints202509.1405.v1

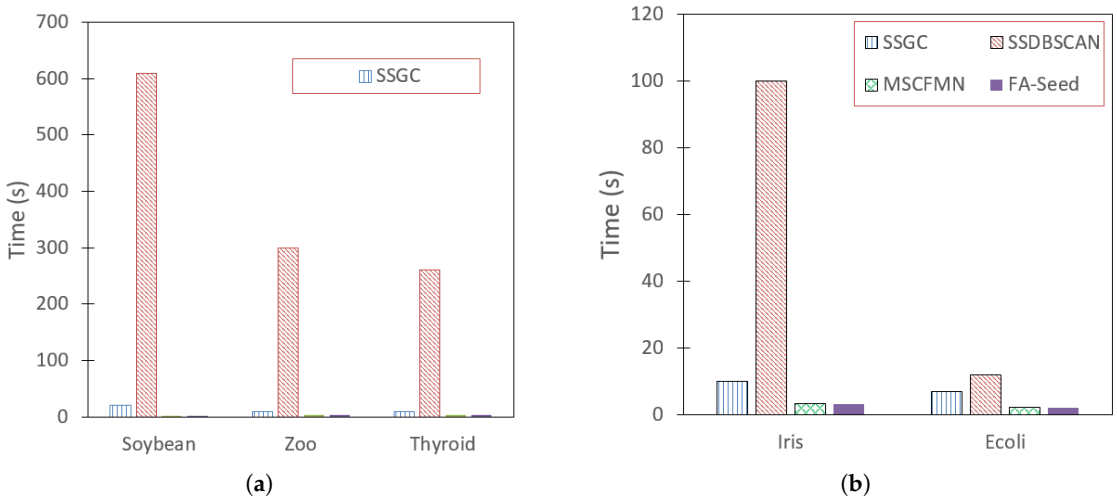


Figure 10. Comparison of Time (measured in Seconds) values between FA-Seed and baseline algorithms on benchmark datasets, including: (**a**) Soybean, Zoo, and Thyroid datasets, (**b**) Iris, and Ecoli datasets.

5. Conclusions

In this paper, we propose FA-Seed, a flexible and active learning-based seed selection model for semi-supervised clustering. By integrating adaptive hyperbox-based data partitioning with an active querying mechanism, FA-Seed effectively identifies high-quality seeds, reduces labeling costs, and improves clustering accuracy. Experimental results on standard benchmark datasets demonstrate that FA-Seed consistently outperforms baseline approaches such as SSGC, MSCFMN, and FC2 in terms of clustering quality, seed efficiency, and computational cost. These findings confirm the potential of FA-Seed for practical applications, particularly in medical diagnosis, where data distributions are complex and labeled data are scarce.

Despite its advantages, FA-Seed still has certain limitations. Its performance is sensitive to parameter settings (e.g., θ) and remains dependent on expert feedback when clusters are highly overlapping or noisy. In the future, we plan to enhance the robustness of FA-Seed against noisy labels and uncertain expert feedback. Combining FA-Seed with deep representation learning may further improve its performance on complex data, while explainable AI techniques could enhance interpretability in sensitive domains such as healthcare. Finally, large-scale validation on multi-domain and multi-center datasets will be essential to confirm the practical applicability of FA-Seed in diagnostic systems and beyond.

Funding: Hanoi University of Industry under Grant 21-2023-RD/HD-DHCN.

Data Availability Statement: The data presented in this study are openly available in the UC Irvine Machine Learning Repository at https://archive.ics.uci.edu, reference number [21], and in the Clustering Datasets at https://cs.joensuu.fi/sipu/datasets/, reference number [16].

Acknowledgments: We would like to express our sincere gratitude to Hanoi University of Industry for providing the support and favorable conditions that enabled us to complete this research.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CS-PDS	Constraints Self-learning from Partial Discriminant Spaces
DPP	Determinantal Point Processes
EM	Expectation-Maximization
FC2	Finding Clusters on Constraints

 doi:10.20944/preprints202509.1405.v1

HX	Hyperbox
HXs	Hyperboxes
FMNN	Fuzzy Min–Max Neural Network
GFMNN	General FMNN
MSCFMN	Modified SCFMN
SCFMN	Semi-supervised Clustering in FMNN
SSGC	Semi-Supervised Graph-based Clustering
SSC	Semi-supervised clustering
SSFC	Semi-supervised Fuzzy Clustering

References

1. Li, F.; Yue, P.; Su, L. Research on the convergence of fuzzy genetic algorithm based on rough classification. In *Proceedings of the International Conference on Natural Computation*; Springer: [Location not specified], 2006; pp. 792–795.
2. Vu, D.M.; Nguyen, V.H.; Le, B.D. Semi-supervised clustering in fuzzy min-max neural network. In *Proceedings of the International Conference on Advances in Information and Communication Technology*; Springer: [Location not specified], 2016; pp. 541–550.
3. Pedrycz, W.; Waletzky, J. Fuzzy clustering with partial supervision. *IEEE Trans. Syst. Man Cybern. Part B (Cybern.)* **1997**, *27*, 787–795.
4. Gabrys, B.; Bargiela, A. General fuzzy min-max neural network for clustering and classification. *IEEE Trans. Neural Netw.* **2000**, *11*, 769–783.
5. Endo, Y.; Hamasuna, Y.; Yamashiro, M.; Miyamoto, S. On semi-supervised fuzzy c-means clustering. In *Proceedings of the 2009 IEEE International Conference on Fuzzy Systems*; IEEE: [Location not specified], 2009; pp. 1119–1124.
6. Minh, V.D.; Ngan, T.T.; Tuan, T.M.; Duong, V.T.; Cuong, N.T. An improvement in integrating clustering method and neural network to extract rules and application in diagnosis support. *Iran. J. Fuzzy Syst.* **2022**, *19*, 147–165.
7. Vu, V.V. An efficient semi-supervised graph based clustering. *Intell. Data Anal.* **2018**, *22*, 297–307.
8. Vu, D.M.; Nguyen, T.S.; Nguyen, V.T.; Nguyen, D.L. FMN-Voting: A Semi-supervised Clustering Technique Based on Voting Methods Using FMM Neural Networks. In *Proceedings of the International Conference on Advances in Information and Communication Technology*; Springer: [Location not specified], 2024; pp. 137–150.
9. Almazroi, A.A.; Atwa, W. Semi-Supervised Clustering Algorithms Through Active Constraints. *Int. J. Adv. Comput. Sci. Appl.* **2024**, *15*, 7.
10. Shen, Z.; Lai, M.J.; Li, S. Graph-based semi-supervised local clustering with few labeled nodes. *arXiv preprint arXiv:2211.11114* **2022**.
11. Tran, T.N.; Vu, D.M.; Tran, M.T.; Le, B.D. The combination of fuzzy min–max neural network and semi-supervised learning in solving liver disease diagnosis support problem. *Arab. J. Sci. Eng.* **2019**, *44*, 2933–2944.
12. Bajpai, N.; Paik, J.H.; Sarkar, S. Balanced seed selection for K-means clustering with determinantal point process. *Pattern Recognit.* **2025**, *164*, 111548.
13. Sun, X. Semi-Supervised Clustering via Constraints Self-Learning. *Mathematics* **2025**, *13*, 1535.
14. Yi, J.; Zhang, L.; Yang, T.; Liu, W.; Wang, J. An efficient semi-supervised clustering algorithm with sequential constraints. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; ACM: [Location not specified], 2015; pp. 1405–1414.
15. Singh, Y.; Susan, S. Lung cancer subtyping from gene expression data using general and enhanced fuzzy min–max neural networks. *Eng. Rep.* **2023**, *5*, e12663.
16. Kämäräinen, J.K.; Kauppi, O.P.; Fränti, P. Clustering Datasets. Available online: https://cs.joensuu.fi/sipu/datasets/ (accessed on 2 August 2025).
17. Khuat, T.T.; Gabrys, B. A comparative study of general fuzzy min-max neural networks for pattern classification problems. *Neurocomputing* **2020**, *386*, 110–125.
18. Vu, D.M.; Nguyen, T.S.; Phung, T.N.; Vu, T.H. Choosing Data Points to Label for Semi-supervised Learning Based on Neighborhood. In *Proceedings of the International Conference on Advances in Information and Communication Technology*; Springer: [Location not specified], 2023; pp. 134–141.
19. Simpson, P.K. Fuzzy min-max neural networks-part 2: Clustering. *IEEE Trans. Fuzzy Syst.* **2002**, *1*, 32.
20. Simpson, P.K. Fuzzy min-max neural networks. I. Classification. *IEEE Trans. Neural Netw.* **1992**, *3*, 776–786.

 doi:10.20944/preprints202509.1405.v1

21. Aha, D. UC Irvine Machine Learning Repository. Available online: https://archive.ics.uci.edu/ (accessed on 2 August 2025).
22. Batista, G.E.A.P.A.; Monard, M.C. An analysis of four missing data treatment methods for supervised learning. *Appl. Artif. Intell.* **2003**, *17*, 519–533.