MINISTRY OF EDUCATION AND TRAINING

VIETNAM ACADEMY OF SIENCE AND TECHNOLOGY

GRADUATE UNIVERSITY OF SIENCE AND TECHNOLOGY

Nguyen Thanh Son

RESEARCH AND DEVELOPMENT OF SEMI-SUPERVISED CLUSTERING MODELS BASED ON FUZZY MIN-MAX NEURAL NETWORKS

SUMMARY OF DISSERTATION ON COMPUTER

Major: Information System

Code : 9 48 01 04

Hanoi - 2026

The dissertation is completed at: Graduate University of Science and Technology, Vietnam Academy of Science and Technology

Supervisors:

Supervisor 1: Assoc. Prof. Dr. Le Ba Dung

Supervisor 2: Dr. Vu Dinh Minh

Referee 1: Assoc. Prof. Dr. Nguyen Van Nui

Referee 2: Assoc. Prof. Dr. Pham Van Hai

Referee 3: Assoc. Prof. Dr. Tran Thi Ngan

Graduate University-level Doctoral Dissertation Evaluation Council at the Graduate University of Science and Technology, Vietnam Academy of Science and Technology, at 2:00 p.m. on September 8, 2026..

This dissertation can be found at:

1) Graduate University of Science and Technology Library

2) National Library of Vietnam

INTRODUCTION

1. Rationale

The rapid growth of data in scale, dimensionality, and complexity has created an increasing need for data-mining methods capable of learning from limited supervisory information. In many real-world problems, unlabeled data can be collected in large quantities, whereas labeling is time-consuming, costly, and often requires expert knowledge. Semi-supervised clustering is therefore an appropriate approach because it combines the intrinsic structure of unlabeled data with a small amount of supervisory information [1, 2].

Fuzzy Min-Max Neural Networks (FMNs) offer advantages such as incremental learning, representation of data by fuzzy hyperboxes, and no requirement to specify the number of clusters in advance. However, semi-supervised FMN models still face difficulties when data distributions are uneven, overlapping, or noisy. In particular, label inference and propagation may depend excessively on a single hyperbox or representative sample. If that information source is unreliable, labeling errors may be amplified during learning [15, 16, 17, 18, 19, 21].

Based on these limitations, the dissertation develops semi-supervised clustering models based on FMNs in two directions: improving the reliability of label inference through majority voting and improving the quality of the seed set through an active selection strategy. The FMN-Voting, FMN-SAL, and FA-Seed models are developed to realize these research directions.

2. Research objectives

The overall objective is to study and develop semi-supervised clustering models based on Fuzzy Min-Max Neural Networks to improve stability, reliability, and learning performance under limited labeled information.

The specific objectives are: (i) to investigate mechanisms exploiting hyperbox structures and neighborhood information for more reliable label inference; (ii) to develop majority-voting mechanisms based on neighboring hyperboxes and neighboring labeled samples; (iii) to construct an active seed-selection strategy considering both representativeness and reliability; and (iv) to experimentally evaluate and compare the proposed models on benchmark and real-world datasets.

3. Research object and scope

The research object comprises semi-supervised clustering models based on FMNs, with emphasis on hyperbox-based data representation, label inference and propagation, majority voting, and seed selection. The experimental scope includes benchmark datasets with different cluster structures and real-world liver-disease data, evaluated using appropriate clustering and classification metrics.

4. Research methodology

The dissertation combines theoretical and experimental research. Based on a review of FMNs and semi-supervised clustering, limitations are analyzed to formulate improved models, mathematical expressions, and algorithms. Computational complexity is analyzed, and experiments are evaluated using Accuracy, Precision, Recall, Specificity, F1-Score, RI, ARI, and NMI.

5. New contributions

First, the dissertation proposes FMN-Voting, which uses majority voting among neighboring hyperboxes to reduce dependence on a single hyperbox in label inference. Second, FMN-SAL changes the voting source from hyperboxes to neighboring labeled samples, thereby reducing the influence of hyperboxes with uneven representativeness or noise. Third, FA-Seed is proposed as an active FMN-based seed-selection strategy that simultaneously considers representativeness and reliability to improve seed-set quality and reduce erroneous label propagation.

6. Dissertation structure

In addition to the Introduction and Conclusion, the dissertation consists of three chapters. Chapter 1 presents an overview of semi-supervised clustering and FMNs. Chapter 2 presents FMN-Voting and FMN-SAL. Chapter 3 presents the FA-Seed model and its experimental evaluation.

CHAPTER 1
OVERVIEW OF SEMI-SUPERVISED CLUSTERING USING FUZZY MIN-MAX NEURAL NETWORKS

1.1. Overview of data clustering and applications

Clustering is an unsupervised learning technique that groups similar objects into the same cluster and separates dissimilar objects into different clusters. In hard clustering, each object belongs to only one cluster; in fuzzy clustering, an object may belong to multiple clusters with different membership degrees. Fuzzy semi-supervised clustering combines the ability of fuzzy clustering to represent uncertainty with a small amount of supervisory information [41, 42].

Representative approaches include seed-based clustering, pairwise constraints, prototype-based models, and fuzzy neural networks. Seed-based methods use a small set of labeled samples to guide clustering, reducing dependence on random initialization and improving stability [43, 44, 45, 46, 47, 48].

1.2. Semi-supervised clustering based on Fuzzy Min-Max Neural Networks

1.2.1. Introduction to Fuzzy Min-Max Neural Networks

FMNs represent data by fuzzy hyperboxes in the feature space. Each hyperbox B_j is determined by a minimum vertex vector V_j and a maximum vertex vector W_j. This structure enables incremental learning, allowing a hyperbox to expand to cover a new sample or creating a new hyperbox when existing hyperboxes cannot be expanded [15, 16, 76].

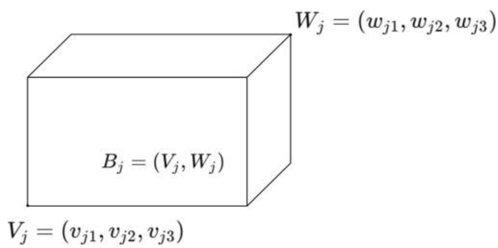


Figure 1.3. *Hyperbox B_j in a three-dimensional space with vertices V_j and W_j.*

1.2.2. Fuzzy hyperbox membership function

The membership degree of sample x_r in hyperbox B_j, denoted by $\mu(x_r, B_j)$, reflects the compatibility between the sample and the region represented by the hyperbox. The fuzziness parameter γ controls the rate at which membership decreases when the sample lies outside the hyperbox boundaries.

$$\mu(x_r, B_j) = \frac{1}{n}\sum_{i=1}^{n}\left[1 - f(x_{ri} - w_{ji}, \gamma) - f(v_{ji} - x_{ri}, \gamma)\right] \quad (1.3)$$

$$f(s,z) = \begin{cases} 1, & s \cdot z > 1 \\ s \cdot z, & 0 \le s \cdot z \le 1 \\ 0, & s \cdot z < 0 \end{cases} \quad (1.4)$$

Hyperbox size is controlled by the maximum limit δ^{max}. When a new sample is considered, the model identifies a suitable hyperbox, checks the expansion condition, updates the vertices, and then checks for overlap.

1.2.3. FMN neural network architecture

The FMN architecture consists of an input layer, a hyperbox layer, and an output layer. The hyperbox layer represents data regions and computes membership degrees; the output layer aggregates information from hyperboxes to determine a class or cluster. In supervised learning, the association between hyperboxes and classes is represented by a binary matrix.

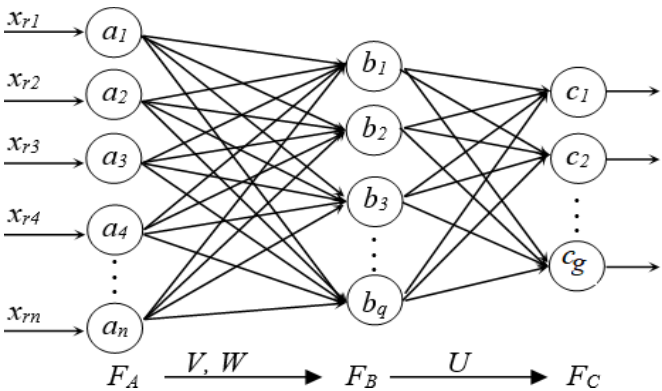


Figure 1.12. *FMN neural network architecture for supervised learning.*

1.2.4. FMN learning algorithm

FMN learning algorithms include supervised, unsupervised, and semi-supervised learning. Learning is based on hyperbox selection, expansion-condition checking, vertex updating, overlap checking, and overlap adjustment. In semi-supervised learning, labeled and unlabeled data are exploited simultaneously [15, 16, 19].

For each sample, if a suitable hyperbox exists and the size condition is satisfied, the hyperbox is expanded. If expansion is not possible, a new hyperbox is created. This incremental learning mechanism allows the FMN to adapt to new data without retraining the entire network.

Algorithm 1.3. Semi-supervised learning algorithm in FMN

Input: $X = (x_r, \ell_{x_r})_{r=1}^{M}$, γ, δ^{max}, β

Output: $\mathcal{B}$

repeat

Select an input sample $(x_r, \ell_{x_r}) \in X$;

if $\exists B_j \in \mathcal{B}$ satisfying $\mu(x_r, B_j) = \max_{B_k \in \mathcal{B}} \mu(x_r, B_k)$ and $\delta(x_r, B_j) \le \delta^{max}$ **then**

Adjust B_j according to (1.6) and (1.7); $X \leftarrow X \setminus (x_r, \ell_{x_r})$;

if overlap is detected between B_j and B_k with $\ell(B_j) \ne \ell(B_k)$ **then**

Adjust the overlap between hyperboxes B_j and B_k;

end if

else

if $\ell(x_r) \ne 0$ **then**

Create a new hyperbox B_{new}; $\ell(B_{new}) \leftarrow \ell_{x_r}$; $\mathcal{B} \leftarrow \mathcal{B} \cup B_{new}$

else

if $\exists B_j \in \mathcal{B}$ such that $\mu(x_r, g(B_j)) > \beta$ **then**

Create a new hyperbox B_{new}; $\ell(B_{new}) \leftarrow \ell(B_j)$; $\mathcal{B} \leftarrow \mathcal{B} \cup B_{new}$;

end if

$\ell_{x_r} = \ell(B_j)$;

end if

16: **end if**
17: **until** $X = \varnothing$
18: **return** $\mathcal{B}$, $(x_r, \ell_{x_r})_{r=1}^{M}$

1.2.5. Issues requiring further research to improve FMN quality for data clustering

FMN quality depends on hyperbox structure and size. When δ^{max} is small, the number of hyperboxes increases; when δ^{max} is large, overlap may increase and discrimination may decrease. Hyperboxes also have different levels of representativeness and reliability, although many models still treat them as equivalent in label propagation [17, 18, 19].

Under limited labeled information, seed selection and the source used for label inference become particularly important. Seeds located in boundary or noisy regions may propagate incorrect labels; similarly, an unstable hyperbox may become an unreliable supervisory source.

1.3. Limitations, research problems, and research directions

The review identifies three main problems in semi-supervised FMN models. First, label inference based on a single entity may be unstable. Second, neighboring hyperboxes differ in quality; therefore, direct voting from hyperboxes may still be affected by noise and uneven data distributions. Third, random or non-adaptive seed selection cannot guarantee representativeness under a limited labeling budget [17, 18, 19, 21].

Accordingly, the dissertation develops: (i) a majority-voting mechanism using multiple hyperboxes to improve label-inference stability; (ii) a voting mechanism based on neighboring labeled samples to reduce the influence of poorly representative hyperboxes; and (iii) active seed selection based on FMN structure while considering representativeness and reliability.

1.4. Chapter 1 conclusion

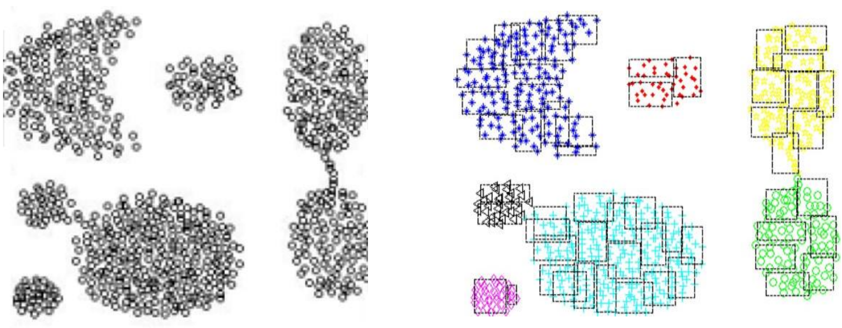
Chapter 1 systematizes the theoretical foundations of semi-supervised clustering and FMNs, clarifying hyperbox representation, membership functions, network architecture, and learning mechanisms. The identified limitations provide the direct basis for FMN-Voting, FMN-SAL, and FA-Seed in the subsequent chapters. Part of Chapter 1 was published in [CT1].

CHAPTER 2
SEMI-SUPERVISED CLUSTERING BASED ON FUZZY MIN-MAX NEURAL NETWORKS WITH MAJORITY VOTING

2.1. Semi-supervised learning based on fuzzy hyperboxes

2.1.1. Data-space partitioning using FMN

FMNs partition the data space into hyperboxes. These hyperboxes serve both as units representing data structure and as the basis for label inference. This representation reduces the number of objects to be processed compared with operating directly on all samples [15, 16, 76].



(a) Data distribution ***(b) Hyperbox partitioning***

Figure 2.1. *Hyperbox partitioning of the Aggregation dataset.*

2.1.2. Hyperbox evaluation based on density and within-hyperbox data distribution

Not all hyperboxes have the same quality. The density and distribution of samples within a hyperbox provide information about its representativeness and stability. Hyperbox evaluation therefore provides a basis for controlling the information sources used in semi-supervised learning.

2.1.3. Semi-supervised mechanism based on hyperboxes

Label information from labeled hyperboxes is used to support the assignment of labels to objects with unknown labels. However, if a decision is based on only one hyperbox, any deviation in that hyperbox can directly affect the result. This motivates the introduction of majority voting into FMNs [19, 21].

2.2. Semi-supervised learning based on majority voting in FMNs

Majority voting aggregates decisions from multiple entities instead of relying on a single source. In FMNs, voting entities may be neighboring hyperboxes or neighboring labeled samples. The label receiving the largest number of votes is preferentially assigned to an unlabeled object [21].

2.3. FMN-Voting model

FMN-Voting infers the label of an unlabeled sample using majority voting among neighboring hyperboxes with high membership degrees. Instead of using a single hyperbox, the model simultaneously exploits multiple hyperboxes, thereby improving decision stability in overlapping or uncertain data regions [21].

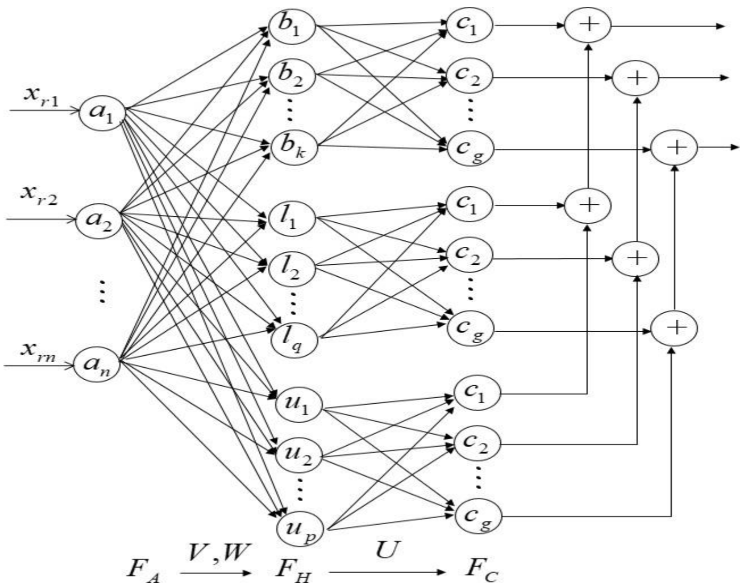


Figure 2.3. *Architecture of the FMN-Voting neural network.*

2.3.3. Learning algorithm in FMN-Voting

The FMN-Voting learning algorithm consists of two stages. Stage 1 constructs a set of labeled seed hyperboxes from the labeled sample set X_L. For each $x_r \in X_L$, a same-label hyperbox B_j with maximum membership $\mu(x_r, B_j)$ is selected, subject to the hyperbox-size condition $\delta(x_r, B_j) \leq \delta^{max}$

. If the condition is satisfied, B_j is expanded; otherwise, a new hyperbox B_{new} is created and assigned the label of x_r.

Stage 2 performs semi-supervised learning by majority voting. For each unlabeled sample $x_r \in X_U$, the k neighboring hyperboxes with the highest membership values are identified. Its label is inferred according to the majority-voting rule:

$$\hat{t}(x_r) = \arg\max_c \sum_{B_j \in N_k(x_r)} I(t(B_j) = c)$$

where $N_k(x_r)$ is the set of k neighboring hyperboxes of x_r. After $\hat{t}(x_r)$ is determined, the sample is used to expand a suitable hyperbox or to create a new hyperbox if the condition $\delta(x_r, B_j) \leq \delta^{max}$ is not satisfied.

2.3.4. Computational complexity of FMN-Voting

Let $M = |X|$ be the number of samples, n the number of attributes, and $K = |B|$ the number of hyperboxes. The processing cost for each sample is bounded by $O(nK)$. Therefore, the semi-supervised training stage has complexity $O(nMK)$. Since both main stages have the same order of complexity, the total complexity remains:

$$T = T_1 + T_2 = O(nMK) + O(MnK) = O(MnK) \quad (2.22)$$

2.4. FMN-SAL model

FMN-Voting assumes that neighboring hyperboxes have equivalent levels of representativeness. This assumption does not always hold when data within hyperboxes are unevenly distributed or noisy. FMN-SAL addresses this limitation by directly exploiting neighboring labeled samples for voting, thereby reducing dependence on the geometric quality of individual hyperboxes [19, 21].

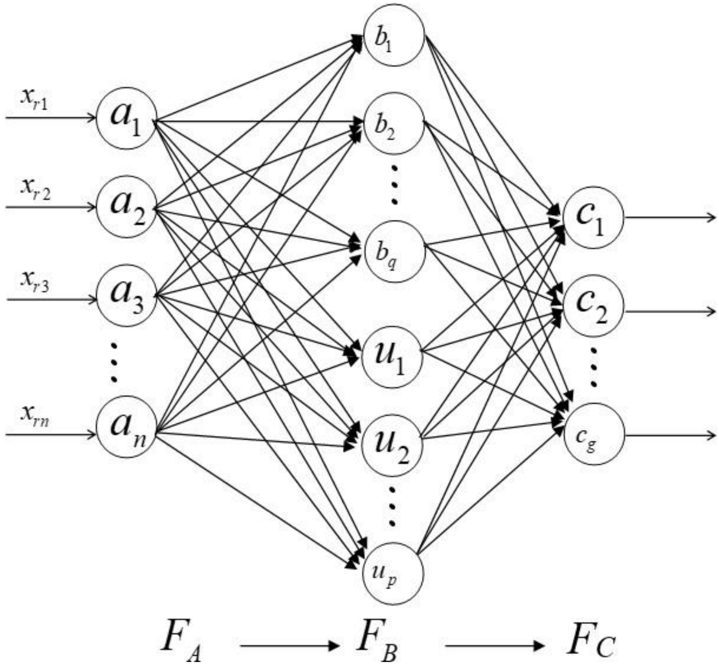


Figure 2.5. *Architecture of the FMN-SAL neural network.*

2.4.3. Learning algorithm in FMN-SAL

The algorithm partitions the data using FMN, maintains the labeled samples, and, when label inference is required for an unlabeled object, identifies an appropriate neighborhood and then performs voting based on the labels of neighboring samples. This mechanism aims to limit the influence of imbalanced hyperboxes or hyperboxes that do not adequately represent the data distribution.

The FMN-SAL algorithm performs semi-supervised learning on the dataset X under two main cases. For each sample x_r, the algorithm selects the hyperbox B_j with the highest membership value satisfying the expansion condition:

$$\mu(x_r, B_j) = \max_{B_k \in B} \mu(x_r, B_k) \text{ and } \delta(x_r, B_j) \le \delta^{max}.$$

If a suitable B_j exists, the hyperbox is adjusted according to (1.6) and (1.7). For an unlabeled sample with $\ell_{x_r} = 0$, the sample is temporarily assigned the label of the hyperbox, $\ell_{x_r} \leftarrow \ell(B_j)$. Subsequently, overlapping hyperboxes with different labels are detected and adjusted. If no suitable hyperbox exists, for $\ell_{x_r} \neq 0$, the algorithm creates a new hyperbox B_{new} and assigns $\ell(B_{new}) \leftarrow \ell_{x_r}$. Conversely, for an unlabeled sample, the algorithm determines the neighborhood set $N(x_r)$ according to (2.23) and performs voting based on neighboring labeled samples. When $|N(x_r)| = k$, the majority label is assigned to x_r according to (2.25).

The process is repeated until $X = \varnothing$; subsequently, $X \leftarrow X'$ is updated, and U is updated based on the hyperbox set B according to (1.36). The essential characteristic of FMN-SAL is that neighboring labeled samples, rather than neighboring hyperboxes, are used for label-inference voting, thereby reducing the influence of hyperboxes with uneven levels of representativeness.

2.5. Experiments and Evaluation

2.5.1. Experimental Setup

Experimental objectives. The experiments were conducted to evaluate the effectiveness of the two proposed models, FMN-Voting and FMN-SAL, focusing on their clustering performance under low proportions of labeled data, their stability under parameter variations, and their comparative performance against baseline methods on datasets with different structures.

Experimental methodology. The models were evaluated on the benchmark datasets Aggregation, Jain, R15, and Flame, as well as the classification datasets Ecoli, Soybean, Zoo, and Yeast. Experiments were also conducted on the ILPD dataset. Model performance was evaluated using RI, ARI, and NMI for clustering tasks, and Accuracy, Precision, Recall, Specificity, and F1-Score for labeled-data tasks. The results of FMN-Voting and FMN-SAL were compared with FMN-based models and the baseline methods considered in the dissertation.

Parameter settings. The main parameters include the fuzziness parameter γ, the maximum hyperbox size δ^{max}, the confidence threshold β, and the number of neighboring objects k used for voting. The experiments varied δ^{max} and the proportion of labeled samples to investigate their effects on the number of hyperboxes and clustering performance. The remaining parameters were configured according to the specific experimental scenarios defined in the dissertation.

2.5.2. Experimental Results and Evaluation

On the benchmark datasets, FMN-SAL achieved high and stable ARI values. Specifically, on Aggregation, Jain, R15, and Flame, the ARI values obtained by FMN-SAL were approximately 0.9911, 0.9879, 0.9980, and 0.9798, respectively. These results indicate that the mechanism of exploiting neighboring labeled samples is suitable for datasets with complex cluster structures.

Table. Summary of representative ARI results (extracted from Table 2.4 of the dissertation).

Model	*Aggregation*	*Jain*	*R15*	*Flame*
GCAE	0.992	1	0.9928	0.9501
DBSCAN	0.9664	0.9382	0.9375	0.9388
FMN-Voting	-	-	0.97804	0.96218
FMN-SAL	0.991139	0.987856	0.998	0.979817

On ILPD, FMN-SAL achieves Accuracy 0.8806 ± 0.0043, Specificity 0.8001 ± 0.0082, Precision 0.8786 ± 0.0044, and F1-Score 0.8788 ± 0.0043. FMN-Voting achieves Accuracy 0.8045 ± 0.0366, higher than SCFMN; K-means++, EM, and DBSCAN obtain lower Accuracy than the proposed FMN models.

Table 2.12. *Comparison of FMN-SAL, FMN-Voting, and benchmark methods.*

Model	**Accuracy**	**Recall**	**Specificity**	**FP Rate**	**Precision**	**F1-Score**
K-means++	0.6252 ± 0.0222	0.6252 ± 0.0222	0.4463 ± 0.0932	0.5537 ± 0.0932	0.6148 ± 0.0437	0.6154 ± 0.0240
EM	0.5647 ±0.0351	0.5647 ±0.0351	0.7353 ±0.1299	0.2647 ±0.1299	0.7589 ±0.0544	0.5648 ±0.0249
DBSCAN	0.6416 ± 0.0022	0.6416 ± 0.0022	0.4411 ± 0.0048	0.5589 ± 0.0048	0.6229 ± 0.0021	0.6306 ± 0.0019
SCFMN	0.7780 ± 0.0184	0.7780 ± 0.0184	0.6442 ± 0.0330	0.3558 ± 0.0330	0.7718 ± 0.0185	0.7736 ± 0.0186
FMN-SAL	0.8806 ± 0.0043	0.8806 ± 0.0043	0.8001 ± 0.0082	0.1999 ± 0.0082	0.8786 ± 0.0044	0.8788 ± 0.0043
FMN-Voting	0.8045	0.8045	0.7029	0.2971	0.8021	0.8030

	± 0.0366	± 0.0366	± 0.0490	± 0.0490	± 0.0365	± 0.0364

Chapter 2 conclusion

Chapter 2 develops FMN-Voting and FMN-SAL to improve the reliability of label inference through voting. FMN-Voting aggregates labels from neighboring hyperboxes, whereas FMN-SAL exploits neighboring labeled samples to reduce the influence of poorly representative hyperboxes. Experimental results show improvements in clustering and classification quality, with FMN-SAL demonstrating stability and generalization across multiple data structures. Part of Chapter 2 was published in [CT2], [CT3], [CT4], and [CT8].

CHAPTER 3

SEMI-SUPERVISED LEARNING IN FMN WITH AN ACTIVE SEED-SELECTION STRATEGY

3.1. Introduction to the FA-Seed model

FMN-Voting and FMN-SAL improve label inference, but learning performance still depends significantly on the quality of the initial seeds. In complex data spaces, seeds do not play equivalent roles; samples in boundary, noisy, or overlapping regions may propagate labeling errors. FA-Seed is developed as a flexible active seed-selection strategy that prioritizes objects with high representativeness and reliability. Its objective is to use a limited labeling budget efficiently and create a seed set capable of effectively guiding semi-supervised clustering [102].

3.2. FA-Seed: Active seed selection based on FMN

Given a dataset X consisting of samples in an n-dimensional space, only a small proportion can be queried/labeled. The problem is to select appropriate samples or hyperboxes to form a seed set and then use this set for semi-supervised learning.

FA-Seed uses the FMN hyperbox structure to organize data and evaluate seed candidates. Instead of random selection, the model considers the representativeness and reliability of each seed. The selection process is integrated with active learning: informative candidates are identified, labels are queried within a limited budget, and the seed set is updated [102].

Hyperbox eccentricity is defined as:

$$e(B_j) = \| g(B_j) - d(B_j) \|_2 \tag{3.8}$$

where $g(B_j)$ is the geometric center and $d(B_j)$ is the data center of the hyperbox. A smaller $e(B_j)$ indicates a more balanced distribution of data within the hyperbox.

The extended seed set from selected hyperboxes is:

$$S = \bigcup_{j \in J_Q} \{x_r \in X |\ \mu(x_r, B_j) \geq \tau\}, \quad \tau \in [0,1] \tag{3.9}$$

The core set of hyperbox B_j is:

$$T_j = \{r \mid \mu(x_r, B_j) \geq \tau\} \quad (3.10)$$

The index set of pure hyperboxes is:

$$J_Q = \{j \mid B_j \in B, e(B_j) \leq \varepsilon\} \quad (3.11)$$

The candidate seed set is:

$$Q = \{d(B_j) \mid j \in J_Q\} \quad (3.12)$$

3.2.3. FA-Seed neural network architecture

The FA-Seed architecture inherits the FMN hyperbox representation layer and adds mechanisms for evaluating and selecting seeds for semi-supervised learning. Separating hyperbox-structure construction from supervisory-source selection enables better control over the quality of label information.

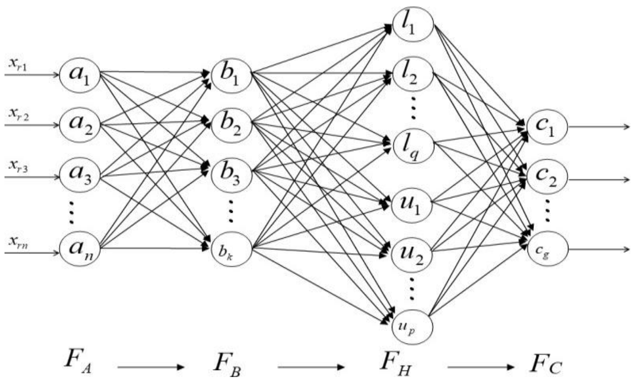


Figure 3.9. *Architecture of the FA-Seed neural network.*

3.2.4. Learning algorithm of the FA-Seed model

The FA-Seed algorithm consists of the following main steps: partitioning data using FMN; evaluating hyperboxes/candidates; actively selecting seeds according to representativeness and reliability; querying or receiving labels; and then performing semi-supervised learning and updating the model structure. The adaptive mechanism allows the model to adjust the seed set according to the state of the data rather than fixing it from the beginning.

Algorithm 3.2. Guided Semi-Supervised Training

Input: X ; $\mathcal{B}$; β ; δ^{max}

Output: $\mathcal{B}$; U ; X

$\mathcal{B}lab = \emptyset$; $\mathcal{B}unl = \emptyset$;
for each $x_r \in X$ **do**
For $B_j \in \mathcal{B}lab \cup \mathcal{B}unl$, the index j^* is determined according to (3.19);
if $x_r \subseteq B_{j^*}$ **then**
Expand hyperbox B_{j^*} according to (3.21);
if (B_{j^*} *overlaps with* B_k) *with* $\ell(B_{j^*}) \neq \ell(B_k)$ **then**
Adjust B_{j^*} and B_k according to (1.15)–(1.26);
end if
else
Create B_{new}, and initialize V_{new} and W_{new} using x_r ;
if $\ell_{x_r} \neq 0$ **then**
$\ell(B_{new}) = \ell_{x_r}$; $\mathcal{B}lab = \mathcal{B}lab \cup B_{new}$;
else
$\mathcal{B}unl = \mathcal{B}unl \cup B_{new}$;
end if
end if
end for
repeat
for each $B_j \in \mathcal{B}_{unl}$ **do**
Compute the centroid $g(B_j)$ according to (1.40);
if ($\exists B_k^*$ with $B_k \in \mathcal{B}_{lab}$ satisfying (3.22)) **then**
$\ell(B_j) = \ell(B_k)$; $\mathcal{B}lab = \mathcal{B}lab \cup B_j$; $\mathcal{B}unl = \mathcal{B}unl \setminus B_j$;
end if
end for
Decrease β ;
until $\mathcal{B}_{unl} = \emptyset$ **or** $\beta \leq \text{confidence threshold}$;
if $\mathcal{B}_{unl} \neq \emptyset$ **then**
for each $B_j \in \mathcal{B}_{unl}$ **do**
Compute the centroid $g(B_j)$ according to (1.40);

30: Set $k^{\leftarrow} \arg\max_k \mu(g(B_j), B_k)$ *with* $B_k \in \mathcal{B}lab$; $\ell(B_j) = \ell(B_{k^*})$;

31: **end for**

32: **end if**

33: $\mathcal{B} = \mathcal{B}lab \cup \mathcal{B}unl$;

34: **return** $\mathcal{B}$.

The computational complexity of FA-Seed is:

$$T = T_1 + T_2 = O(nMK) + O(nMK) = O(nMK). \qquad (3.27)$$

where $M = |X|$ is the number of samples, n is the number of attributes, and $K = |B|$ is the number of hyperboxes. In practice, K is typically much smaller than M, allowing FA-Seed to scale as the dataset size increases.

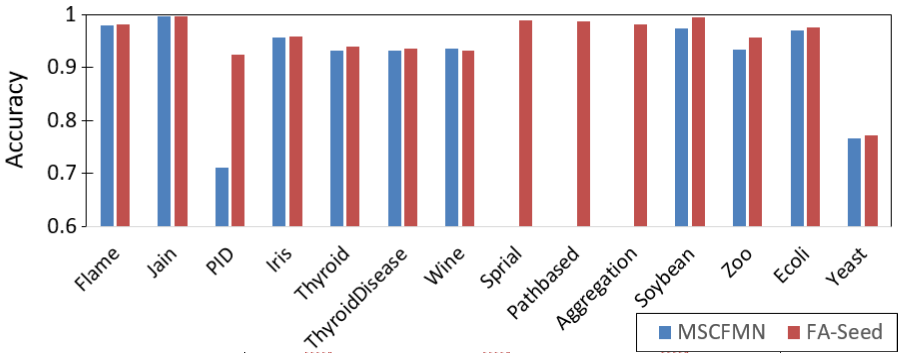
3.3. Experiments and evaluation

FA-Seed is evaluated on benchmark datasets and liver-disease data and compared with MSCFMN and other clustering methods. The experiments examine the influence of δ^{max}, class-recognition capability, partition quality, and runtime [102].

On the liver-disease dataset, FA-Seed achieves RI 0.8215 ± 0.0263, ARI 0.6285 ± 0.0549, and NMI 0.4779 ± 0.0577, outperforming MSCFMN and the K-means++, EM, and DBSCAN methods in the reported experiments.

Model	**RI**	**ARI**	**NMI**
K-means++	0.5314 ± 0.0111	0.0189 ± 0.0336	0.0107 ± 0.0110
EM	0.5097 ± 0.0146	−0.0130 ± 0.0144	0.0926 ± 0.0343
DBSCAN	0.5393 ± 0.0013	0.0347 ± 0.0030	0.0066 ± 0.0011
MSCFMN	0.6783 ± 0.0402	0.3279 ± 0.0812	0.1993 ± 0.0665
FA-Seed	0.8215 ± 0.0263	0.6285 ± 0.0549	0.4779 ± 0.0577

Runtime analysis shows that FA-Seed has an advantage over baseline methods in the reported experiments. With complexity $O(nMK)$, the model exploits the fact that the number of hyperboxes K is generally much smaller than the number of samples M. In contrast, graph-based methods may incur high costs when neighborhood relations are constructed for large-scale data [103].



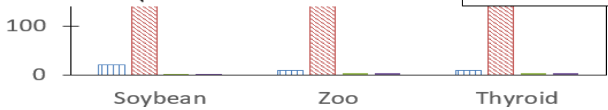


Figure 3.14. *Visual comparison of Accuracy between FA-Seed and MSCFMN.*

Figure 3.15. *Runtime comparison (s) between FA-Seed and baseline algorithms on the experimental datasets.*

Chapter 3 conclusion

Chapter 3 presents FA-Seed with an emphasis on active seed selection. The model does not assume that all seeds play equivalent roles; instead, it simultaneously considers representativeness and reliability. Adaptive selection helps reduce erroneous label propagation in overlapping, noisy, or unevenly distributed regions. Complexity analysis and experiments indicate that FA-Seed is feasible and has the potential to improve clustering quality. Part of Chapter 3 was published in [CT5], [CT6], and [CT7].

CONCLUSION

A. Main results of the dissertation

The dissertation focuses on semi-supervised clustering in FMNs in the context of increasingly large and complex data with scarce supervisory information. Based on a systematic review, it identifies limitations of semi-supervised FMN models when label inference and propagation depend on single entities or when seed quality is not controlled.

The first result is FMN-Voting, which introduces majority voting based on neighboring hyperboxes into label inference. The second result is FMN-SAL, which directly exploits neighboring labeled samples for voting. The third result is FA-Seed, which develops an active FMN-based seed-selection strategy considering both representativeness and reliability.

B. Novel contributions

The dissertation extends semi-supervised FMNs toward greater stability and reliability: from aggregating information from multiple hyperboxes, to directly exploiting neighboring labeled samples, and finally to actively controlling the quality of supervisory information through seed selection.

C. Limitations and future research

Although the proposed models achieve positive results, handling data with complex feature structures and very large scale still requires further research. Model effectiveness also depends on feature representation, hyperbox structure, and label-inference mechanisms.

Future directions include: (i) combining semi-supervised FMNs with deep-learning or feature-representation models for complex data; (ii) developing more adaptive label-inference mechanisms based on reliability and hyperbox dynamics; (iii) investigating parallel and distributed strategies to improve scalability on large datasets; and (iv) further exploiting fuzzy-rule extraction to enhance interpretability and applicability in decision-support systems.

Overall, the dissertation contributes new methodological and application-oriented results to semi-supervised clustering in FMNs, advancing this research direction toward greater stability, reliability, and suitability for real-world data under limited labeled information.

LIST OF PUBLICATIONS RELATED TO THE DISSERTATION

CT1 Nguyen Thanh Son, Vu Thi Duong, Nguyen Anh Thai, Hoang Quang Huy, Vu Dinh Minh. “An Overview of Fuzzy Min-Max Neural Networks and Applications”, Selected Problems of Information and Communication Technology, 25th National Conference, Hanoi, 8–9 December 2022, Science and Technics Publishing House, 2022, pp. 238–244.

CT2 Dinh-Minh Vu, Thanh-Son Nguyen, Trung-Nghia Phung, Trinh-Hoang Vu. “Choosing Data Points to Label for Semi-supervised Learning Based on Neighborhood”, Advances in Information and Communication Technology: Proceedings of ICTA 2023, Lecture Notes in Networks and Systems, vol. 847, Springer, 2023, pp. 134–141.

CT3 Dinh-Minh Vu, Thanh-Son Nguyen, Van Tinh Nguyen, Duc-Luu Nguyen. “FMN-Voting: A Semi-supervised Clustering Technique Based on Voting Methods Using FMM Neural Networks”, Advances in Information and Communication Technology: Proceedings of the 3rd International Conference ICTA 2024, Lecture Notes in Networks and Systems, vol. 1205, Springer Cham, 2025, pp. 137–150. DOI: 10.1007/978-3-031-80943-9_15.

CT4 Thanh-Son Nguyen, Dinh-Minh Vu, Duc-Luu Nguyen. “FMN-SSL: A Semi-supervised Approach for Lung Tumor Segmentation”, technical program ICTA 2025; proceedings Advances in Information and Communication Technology: Proceedings of the 4th International Conference, ICTA 2025, Volume 3, Lecture Notes in Networks and Systems, vol. 1833, Springer Cham, online publication March 2026.

CT5 Dinh-Minh Vu, Thanh-Son Nguyen, Trinh-Hoang Vu, Thi-Duong Vu, Duc-Luu Nguyen, Quang-Huy Hoang, Ba-Dung Le. “FMN-OE: Evaluation and Elimination of Hyperbox Overlap”, Intelligent Systems and Data Science: Third International Conference, ISDS 2025, Proceedings, Part II, Communications in Computer and Information Science, vol. 2714, Springer, Singapore, 2026, pp. 328–342. DOI: 10.1007/978-981-95-3358-9_28.

CT6 Dinh-Minh Vu, Thanh-Son Nguyen. “FA-Seed: Flexible and Active Learning-Based Seed Selection”. Information, 2025, 16, 884. DOI: 10.3390/info16100884. (ESCI, Scopus).

CT7 Nguyen Thanh Son, Vu Dinh Minh, Le Ba Dung, Nguyen Van Thanh. “Semi-supervised Clustering with a Hyperbox-Weight-Based Seed Selection Method in Fuzzy Min-Max Neural Networks”, Selected Problems of

Information and Communication Technology, 24th National Conference, Thai Nguyen, 13–14 December 2021, Science and Technics Publishing House, 2021, pp. 433–438.

CT8 Nguyen Thanh Son, Vu Dinh Minh. “Integrating Active Learning And Controlled Label Propagation In Fuzzy Min-Max Neural Networks”, Journal of Computer Science and Cybernetics. The manuscript has completed the second round of peer review.