MINISTRY OF EDUCATION AND TRAINING

VIETNAM ACADEMY OF SCIENCE AND TECHNOLOGY

GRADUATE UNIVERSITY OF SCIENCE AND TECHNOLOGY



Nguyen Thanh Son

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

DOCTORAL DISSERTATION SUMMARY

Major: Information Systems

Code: 9 48 01 04

Hanoi - 2026

This dissertation was completed at:

Graduate University of Science and Technology, Vietnam Academy of Science and Technology

Supervisors

1. Associate Professor Dr. Le Ba Dung
   Institute of Information Technology, Vietnam Academy of Science and Technology
2. Dr. Vu Dinh Minh
   School of Information and Communication Technology, Hanoi University of Industry

Reviewer 1: Assoc. Prof. Dr. Nguyen Van Nui, Thai Nguyen University

Reviewer 2: Assoc. Prof. Dr. Pham Van Hai, Hanoi University of Science and Technology
Reviewer 3: Assoc. Prof. Dr. Tran Thi Ngan, Vietnam National University, Hanoi

The dissertation will be defended before the Doctoral Dissertation Evaluation Committee of the Graduate University of Science and Technology, Vietnam Academy of Science and Technology, at the Graduate University of Science and Technology, Vietnam Academy of Science and Technology, at 14 hours 00 minutes, on 08/9/2026

The dissertation is available at:

1. Library of the Graduate University of Science and Technology
2. National Library of Vietnam

INTRODUCTION

1. Research motivation

In recent years, semi-supervised clustering has attracted increasing attention, driven by the need to exploit large and complex datasets when labelled information is scarce and labelling is costly. Combining labelled and unlabelled data allows a model to better exploit the latent structure of the data and thus to improve clustering quality compared with purely unsupervised methods. However, existing methods often struggle when the data are noisy, unevenly distributed or have ambiguous cluster boundaries, or when interpretability is essential, as in decision support and biomedical applications.

Fuzzy clustering lets each object belong to several clusters with different membership degrees and suits data with overlapping regions. The Fuzzy Min-Max Neural Network (FMN) represents data by fuzzy hyperboxes, learns incrementally and yields interpretable IF-THEN rules, and has been extended to semi-supervised learning (GFMM, RFMN, SS-FMM, SSL-FMM, SCFMN, MSCFMN). Nevertheless, these models still have three main limitations:

- First, clustering depends heavily on the quality of the initially labelled samples or hyperboxes, whose representativeness and reliability have not been adequately assessed.

- Second, label inference relies mainly on neighbourhood (membership) relations between hyperboxes and does not sufficiently consider the purity, data density and representative role of each hyperbox, which increases the risk of error propagation.

- Third, the models have difficulty with samples in boundary or overlapping regions, where high uncertainty affects label inference and propagation.

Motivated by these issues, the dissertation investigates new approaches to semi-supervised clustering based on fuzzy min-max neural networks, aiming to improve label inference, reduce error propagation and enhance clustering quality under limited supervision.

2. Research objectives

The general objective is to propose and develop semi-supervised clustering models based on fuzzy min-max neural networks that increase

the reliability of label inference, limit error propagation and improve clustering quality under limited supervision. The specific objectives are:

- To study mechanisms for assessing the quality of initially labelled samples or hyperboxes and to select highly representative and reliable objects as sources of supervision.

- To propose new label inference mechanisms for unlabelled hyperboxes that also take into account purity, data density and the representative role of hyperboxes.

- To build sample-based and/or hyperbox-based label voting strategies that stabilise labelling decisions for data with uneven density, imbalanced clusters or complex boundaries.

- To evaluate the proposed models on benchmark and real-world datasets using appropriate metrics.

3. Research object, scope and content

The research object is semi-supervised clustering based on FMN, focusing on (i) the formation and adjustment of fuzzy hyperboxes during learning, (ii) label inference/propagation under label scarcity, and (iii) seed selection strategies that guide semi-supervised learning. The dissertation addresses three topics: label inference by voting among neighbouring hyperboxes (FMN-Voting); label inference based on neighbouring labelled samples (FMN-SAL); and active/adaptive seed selection (FA-Seed). The evaluation covers benchmark datasets from the UCI and CS repositories and the ILPD medical dataset for liver disease diagnosis.

4. Research methodology

The dissertation combines theoretical analysis with experimental evaluation. Based on a review of semi-supervised clustering and fuzzy min-max neural networks, improvements to label inference and seed selection are proposed; the proposed models are implemented, tested and evaluated on selected datasets using appropriate metrics.

5. Scientific and practical significance

Scientifically, the dissertation incorporates structural information of hyperboxes into label inference and seed selection in semi-supervised FMN. Practically, validation on the ILPD blood test dataset demonstrates

the applicability of the models to diagnosis support when labelled data are limited.

6. New contributions

- The FMN-Voting model for semi-supervised clustering based on FMN, using a majority voting mechanism to stabilise label inference.

- The FMN-SAL model with a voting strategy based on the *k* nearest labelled samples, reducing the influence of imbalanced hyperboxes, hyperboxes with large centre deviation or low usage index during label propagation.

- The FA-Seed model with an active/adaptive seed selection strategy based on hyperbox partitioning and reliability assessment, improving the quality of the seed set and the effectiveness of semi-supervised learning under limited labels.

- The design and execution of experiments on benchmark datasets and a real medical dataset of blood biochemical indicators to evaluate the performance, stability, ability to exploit sparsely labelled data and computational cost of the proposed models.

7. Organisation of the dissertation

Besides the Introduction and Conclusion, the dissertation consists of three chapters. Chapter 1 reviews semi-supervised clustering and fuzzy min-max neural networks, analyses open issues and defines the research direction. Chapter 2 presents the FMN-Voting and FMN-SAL models, their algorithms, complexity analysis and experimental evaluation. Chapter 3 presents the FA-Seed model with a seed selection strategy based on hyperbox structure combined with expert feedback, together with its experimental evaluation.

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

1.1. Data clustering and its applications

Clustering is a key data mining technique within unsupervised learning that groups similar objects into the same cluster. Crisp clustering assigns each point to exactly one cluster, whereas fuzzy clustering allows a point to belong to several clusters with membership values in [0, 1], which suits

data with unclear or overlapping boundaries. Semi-supervised fuzzy clustering extends fuzzy clustering by exploiting partial prior knowledge (labels, must-link/cannot-link constraints or predefined memberships) to guide the clustering process.

Typical approaches are seed-based, pairwise constraint-based, prototype-based and fuzzy neural network-based semi-supervised fuzzy clustering. Seed-based methods depend strongly on seed quality, constraint-based methods face cost and consistency issues, and prototype-based methods depend on prototype quality. FMN flexibly models core, boundary and overlapping regions with fuzzy hyperboxes, learns incrementally and is interpretable; it is therefore chosen as the foundation for FMN-Voting, FMN-SAL and FA-Seed.

1.2. Semi-supervised clustering based on fuzzy min-max neural networks

1.2.1. Fuzzy hyperboxes and membership function

FMN was proposed by Simpson for classification (1992) and extended to clustering (1993). Knowledge in FMN is represented by fuzzy hyperboxes; each hyperbox B_j is defined by a lower bound vector V_j and an upper bound vector W_j:

$$B_j = \{x \in \mathbb{R}^n \mid v_{ji} \le x_i \le w_{ji},\ \forall i = 1, \ldots, n\} \tag{1.2}$$

The membership degree of a data point x_r in hyperbox B_j is computed as

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

where γ is the sensitivity parameter controlling how fast the membership decreases outside the hyperbox, and $f(\cdot,\cdot)$ is the two-parameter ramp function:

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

1.2.2. Data partitioning with fuzzy hyperboxes

Training consists of three basic operations: expansion, overlap test and contraction. When B_j is expanded to cover x_r, its bounds are updated as $v_{ji} \leftarrow \min(v_{ji}, x_{ri})$ and $w_{ji} \leftarrow \max(w_{ji}, x_{ri})$; expansion is accepted only if the hyperbox size does not exceed δ^{max}. If two hyperboxes of different classes overlap, FMN considers four basic overlap cases and adjusts the bounds on the overlapping dimension with minimal change. Sample assignment follows the containment rule ($\mu(x, B_j) = 1$), the expansion rule and the creation of a new hyperbox $B_{new} = (x, x)$.

1.2.3. Network structure and learning algorithms

For unsupervised learning, FMN has two layers: the input layer F_A (n neurons) and the hyperbox layer F_B. For supervised learning, it has three layers F_A, F_B and the output layer F_C; the connection between F_B and F_C is a binary matrix U with $u_{jg} = 1$ if $\ell(B_j) = g$ and 0 otherwise, and the output of class g is $c_g = \max_j(b_j u_{jg})$. The dissertation presents the supervised, unsupervised and semi-supervised FMN learning algorithms; in the semi-supervised one, a new hyperbox created from an unlabelled sample takes the label of the nearest hyperbox satisfying the threshold β.

1.2.4. Open issues

Many FMN improvements have been proposed concerning network structure, learning algorithms, overlap handling and pruning of unreliable hyperboxes, together with the semi-supervised models GFMM, RFMN, SS-FMM, SSL-FMM, SCFMN and MSCFMN. However, existing methods still (i) treat hyperboxes as equally important despite their different quality; (ii) let unreliable hyperboxes take part in label propagation; (iii) handle overlaps with heuristic rules applied uniformly; and (iv) do not separate structure learning from label learning.

1.3. Limitations, research problems and directions

Accordingly, the dissertation aims to assess the quality and reliability of hyperboxes in order to select highly representative hyperboxes for label propagation, to design controlled semi-supervised learning mechanisms, and to separate the structure learning phase from the label learning phase.

1.4. Conclusion of Chapter 1

Chapter 1 presented the theoretical background of semi-supervised fuzzy clustering with a focus on FMN (fuzzy hyperboxes, membership function, network structure and learning algorithms), reviewed the studies from GFMM and RFMN to recent semi-supervised FMN models, and identified limitations in exploiting unlabelled data, assessing reliability and controlling label error propagation. Part of this chapter 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

FMN partitions the data space by a set of fuzzy hyperboxes, so a cluster can be represented by one or more sub-regions; non-convex, arbitrarily shaped and unevenly dense clusters can thus be modelled. To assess the density and distribution of data inside a hyperbox, the dissertation uses the centre deviation $e(B_j)$ between the geometric centre $g(B_j)$ and the data centre $d(B_j)$:

$$e(B_j) = \|g(B_j) - d(B_j)\|_2\,,$$
$$g(B_j) = \frac{V_j + W_j}{2}, \qquad d(B_j) = \frac{1}{|T_j|}\sum_{r\in T_j} x_r \tag{2.1)-(2.2}$$

A hyperbox is considered pure if its centre deviation satisfies condition (2.3); the usage index $c(B_j)$ reflects how much the hyperbox is exploited in representing and classifying data (2.4), where U_j is the proportion of samples of B_j within its group and A_j is the proportion of samples correctly classified by B_j:

$$e(B_j) \le \varepsilon, \quad \varepsilon > 0 \tag{2.3}$$

$$c(B_j) = (1-\varphi)\,U_j + \varphi\,A_j \tag{2.4}$$

In FMN, points lying entirely inside a hyperbox have full membership and are regarded as equivalent, so supervision is raised from data points to hyperboxes: once a representative point is labelled, its label is propagated to all points with full membership in that hyperbox, which reduces labelling

cost and the influence of noise and boundary points. Each labelled hyperbox B_j with $\ell(B_j) = c$ can be converted into a rule R_j: IF $x_1 \in [v_{j1}, w_{j1}]$ AND … AND $x_n \in [v_{jn}, w_{jn}]$ THEN $y = c$, whose reliability is assessed by $|T_j|$, $e(B_j)$ and $c(B_j)$.

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

Sample-based majority voting (e.g. kNN) assumes that nearby samples tend to share the same class; the label of an unknown sample is the dominant label among its k nearest neighbours. The technique is intuitive and adapts well to non-linear distributions, but it depends on data quality, is sensitive to k and becomes unstable when voting directly on individual points in overlapping regions. In FMN, voting can be defined on (1) geometric representatives of hyperboxes - the approach of FMN-Voting [CT3], developed from MSCFMN; or (2) neighbouring labelled samples - the approach of FMN-SAL [CT2], inherited from SS-FMM.

2.3. The FMN-Voting model

2.3.1. Problem description and idea

Given limited labels and a set of hyperboxes formed by FMN, the problem is to exploit the hyperbox structure to derive auxiliary label information and propagate labels to unlabelled regions. Unlike SS-FMM, SCFMN and MSCFMN, which rely on a global quality threshold β and many passes, FMN-Voting infers labels locally: the label of a new hyperbox is decided by majority voting among neighbouring labelled hyperboxes rather than by a single guiding hyperbox. The algorithm has two phases: (1) seed set construction by unsupervised learning, assigning auxiliary labels to samples with full membership; (2) semi-supervised training with hyperbox-based majority voting.

2.3.2. FMN-Voting network architecture

The architecture has three layers F_A, F_H and F_C (Figure 2.3). The hyperbox layer F_H is organised into labelled, intermediate and unlabelled hyperbox groups; the layer F_C has g neurons representing clusters, and signals from several hyperboxes are aggregated by (+) nodes. For an unlabelled sample x_r that cannot be covered or absorbed by a suitable hyperbox, the model determines the neighbour set $N(x_r)$ of labelled

hyperboxes with the highest membership that satisfy the threshold β; each hyperbox in the set casts one vote:

$$\hat{\ell}(x_r) = \arg\max_{c \in \mathcal{C}} \sum_{L_i \in N(x_r)} \mathbb{I}\big(\ell(L_i) = c\big) \tag{2.9}$$

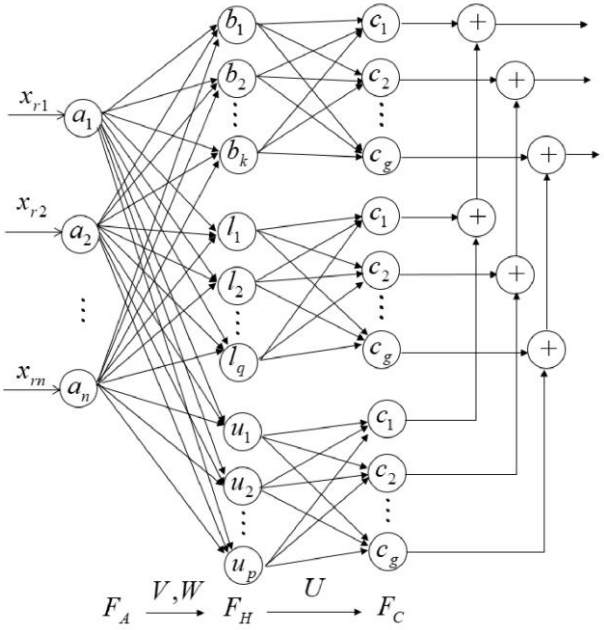


Figure 2.3: *Architecture of the FMN-Voting neural network*

2.3.3. Learning algorithm of FMN-Voting

Phase 1 (Algorithm 2.1) performs unsupervised learning to build the hyperbox set B and assigns auxiliary labels to samples with full membership in a hyperbox; its complexity is $T_1 = O(nMK)$. Phase 2 (Algorithm 2.2) builds the labelled hyperbox set L and infers labels by majority voting over $k = 2C + 1$ neighbouring hyperboxes, where C is the number of classes/clusters; k is odd and larger than $2C$ so that the neighbourhood is large enough and decisions do not rely on too few hyperboxes. When fewer than $2C + 1$ labelled hyperboxes are available, the actual number of neighbours is used:

$$N(x_r) = \underset{\min(2C+1,\,|\mathcal{L}|)}{\arg\operatorname{top}} \big\{\mu(x_r, L_j) \mid L_j \in \mathcal{L},\ \mu(x_r, L_j) \geq \beta\big\} \tag{2.17}$$

$$NoC_\ell = \sum_{L_j \in N(x_r)} \mathbb{I}\big(\ell(L_j) = \ell\big),\ \ \ell = 1, \ldots, C; \tag{2.18}$$

$$\ell(L_{\text{new}}) = \arg\max_{\ell} NoC_\ell \tag{2.19}$$

Algorithm 2.2 Semi-supervised training with hyperbox-based majority voting

Input: $\mathcal{B}$; $X = \{(x_r, \ell_{x_r})\}_{r=1}^{M}$; $\delta_2^{\max}$; γ; β; $R_{\max}$;
Output: $\mathcal{L}$; $X = \{(x_r, \ell_{x_r})\}_{r=1}^{M}$; U;

$\mathcal{L} \leftarrow \{B_j \in \mathcal{B} \mid \ell(B_j) \neq 0\}$; $X^1 \leftarrow \varnothing$; $r \leftarrow 0$;
repeat
  Select an input sample $(x_r, \ell_{x_r}) \in X$;
  if $\exists L_j \in \mathcal{L}$ such that $\mu(x_r, L_j) = 1$ **then**
    if $\ell_{x_r} \neq 0$ **then**
      $\ell_{x_r} \leftarrow \ell(L_j)$; $X^1 \leftarrow X^1 \cup \{(x_r, \ell_{x_r})\}$; $X \leftarrow X \setminus \{(x_r, \ell_{x_r})\}$;
    end if
  else
    if $\exists L_j \in \mathcal{L}$: $\mu(x_r, L_j) = \max_{L_k \in \mathcal{L}} \mu(x_r, L_k)$ and $\delta(x_r, L_j) \leq \delta_2^{\max}$ **then**
      if $\ell_{x_r} \neq 0$ **then**
        $\ell_{x_r} \leftarrow \ell(L_j)$; $X^1 \leftarrow X^1 \cup \{(x_r, \ell_{x_r})\}$; $X \leftarrow X \setminus \{(x_r, \ell_{x_r})\}$;
      end if
      Expand L_j; resolve the overlap between L_j and every $L_k \in \mathcal{L}$ with $\Omega(L_j, L_k) > 0$, $\ell(L_j) \neq \ell(L_k)$;
    else
      Determine the neighbour set $N(x_r)$ by (2.17);
      if $|N(x_r)| \geq 1$ **then**
        Create L_{new} containing x_r; count the votes by (2.18); assign $\ell(L_{\text{new}})$ by (2.19); $\mathcal{L} \leftarrow \mathcal{L} \cup \{L_{\text{new}}\}$; $X^1 \leftarrow X^1 \cup \{(x_r, \ell_{x_r})\}$; $X \leftarrow X \setminus \{(x_r, \ell_{x_r})\}$;
      end if
    end if
  end if
  $r \leftarrow r + 1$;
until $X = \varnothing$ **or** $r = R_{\max}$
$X \leftarrow X^1$; update the class-link matrix U from $\mathcal{L}$ by (1.36);
return $\mathcal{L}$; X; U;

Proposition 2.2 shows that Algorithm 2.2 has complexity $T_2 = O(R_{\max} nMK)$, where $R_{\max}$ is the maximum number of passes of the repeat-until loop. The total cost of FMN-Voting is

$$T = T_1 + T_2 = O(nMK) + O(R_{\max} MnK) = O(R_{\max} nMK) \tag{2.22}$$

FMN-Voting is suitable when neighbouring labelled hyperboxes reflect the local distribution well. However, it treats the selected hyperboxes as equal voters; when data are unevenly distributed inside a hyperbox, its geometric centre may deviate from the actual data centre, lowering voting reliability. Moreover, for non-spherical clusters, hyperbox formation may be sensitive to the order of training samples.

2.4. The FMN-SAL model

2.4.1. Problem description and idea

FMN-SAL directly exploits the labelled samples in the training set to label hyperboxes, forming a set of labelled hyperboxes as the initial supervision source, and then propagates labels to unlabelled hyperboxes and samples through voting. The FMN-Voting assumption that neighbouring hyperboxes are equally representative does not always hold: a hyperbox may be stretched to one side (imbalanced), have a geometric centre deviating from its data centre, or have a low usage index. FMN-SAL therefore keeps the goal of label inference but changes the voting source: neighbouring labelled samples, rather than neighbouring hyperboxes, become the voting units, reducing the dependence on the geometric centre and representativeness of individual hyperboxes.

2.4.2. Architecture and learning algorithm of FMN-SAL

The FMN-SAL architecture consists of the input layer F_A, the hyperbox layer F_B (labelled hyperboxes b_1, …, b_q and hyperboxes whose labels are to be inferred u_1, …, u_p) and the output layer F_C. For a sample x_r, the neighbour set of the k labelled samples with the highest membership is determined by (2.23), and the votes and the label of the new hyperbox by (2.24)-(2.25):

$$N(x_r) = \underset{\min(k,|X'|)}{\arg\operatorname{top}} \left\{ \mu(x_r, x_i) \mid x_i \in X',\ \mu(x_r, x_i) \geq \beta \right\} \tag{2.23}$$

$$NoC_\ell = \sum_{x_i \in N(x_r)} \mathbb{I}\left(\ell_{x_i} = \ell\right); \qquad \ell(B_{\text{new}}) = \arg\max_{\ell} NoC_\ell \tag{2.24)-(2.25}$$

Proposition 2.3: the worst-case time complexity of Algorithm 2.3 is

$$T_{SAL} = O\left(nMK + nM^2\right) = O\left(nM(K+M)\right) \tag{2.26}$$

FMN-SAL reduces the dependence on hyperbox representativeness but depends more on the quality and distribution of labelled samples; hence it does not necessarily outperform other models on every dataset.

Algorithm 2.3 FMN-SAL semi-supervised learning algorithm

Input: $X = \{(x_r, \ell_{x_r})\}_{r=1}^{M}$, $\delta^{\max}$, β
Output: $\mathcal{B}$, $X = \{(x_r, \ell_{x_r})\}_{r=1}^{M}$, U

Initialise the set $\mathcal{U}$ of p hyperboxes by (1.33); $\mathcal{B} \leftarrow \varnothing$; $X' \leftarrow \varnothing$;
Create a hyperbox B_{new} by (1.32); $\mathcal{B} \leftarrow \mathcal{B} \cup \{B_{\text{new}}\}$;
repeat
  Select an input sample $(x_r, \ell_{x_r}) \in X$;
  if $\exists B_j \in \mathcal{B}$: $\mu(x_r, B_j) = \max_{B_k \in \mathcal{B}} \mu(x_r, B_k)$ and $\delta(x_r, B_j) \leq \delta^{\max}$ **then**
    Adjust B_j by (1.6), (1.7); **if** $\ell_{x_r} = 0$ **then** $\ell_{x_r} \leftarrow \ell(B_j)$;
    $X' \leftarrow X' \cup \{(x_r, \ell_{x_r})\}$; $X \leftarrow X \setminus \{(x_r, \ell_{x_r})\}$;
    Resolve the overlap between B_j and B_k if $\ell(B_j) \neq \ell(B_k)$;
  else
    if $\ell_{x_r} \neq 0$ **then**
      Create B_{new}; $\ell(B_{\text{new}}) \leftarrow \ell_{x_r}$; $\mathcal{B} \leftarrow \mathcal{B} \cup \{B_{\text{new}}\}$; $X' \leftarrow X' \cup \{(x_r, \ell_{x_r})\}$; $X \leftarrow X \setminus \{(x_r, \ell_{x_r})\}$;
    else
      Determine the neighbour set $N(x_r)$ by (2.23);
      if $|N(x_r)| = k$ **then**
        Create B_{new}; count the votes by (2.24); assign $\ell(B_{\text{new}})$ by (2.25); $\ell_{x_r} \leftarrow \ell(B_{\text{new}})$; $\mathcal{B} \leftarrow \mathcal{B} \cup \{B_{\text{new}}\}$; $X' \leftarrow X' \cup \{(x_r, \ell_{x_r})\}$; $X \leftarrow X \setminus \{(x_r, \ell_{x_r})\}$;
      end if
    end if
  end if
until $X = \varnothing$
$X \leftarrow X'$; update U from $\mathcal{B}$ by (1.36);
return $\mathcal{B}, X, U$;

2.5. Experiments and evaluation

2.5.1. Experimental setup

The experiments evaluate FMN-Voting and FMN-SAL against FMN variants and related methods, analyse the effect of the label inference strategies, the influence of parameters ($\delta^{\max}$, β, γ) and the computational cost. Ten-fold cross-validation is used; in each fold, the test set is used only for evaluation after training and parameter selection, and preprocessing parameters (including the means used to impute missing values) are estimated from the training set only. Results are reported as mean and standard deviation; no statistical hypothesis tests were performed, so

differences describe experimental trends. Main parameters: $\gamma = 10$, initial $\beta = 0.99$, decay factor $\varphi = 0.9$ and $\beta_{min} \geq 0.8$. Environment: MATLAB R2014a 64-bit, Intel Core i5-6300HQ, 16 GB RAM.

Thirteen datasets are used (Table 2.1 of the dissertation), with 150 to 7200 samples, 2 to 60 attributes and 2 to 15 groups. Aggregation, Flame, Spiral, Jain, R15, Pathbased and Thyroidnew come from the CS repository; Iris, Wine, PID, Thyroid, Sonar and ILPD from the UCI repository. ILPD contains 583 patient records (416 with liver disease and 167 without) described by 10 clinical and blood biochemical attributes. Since the models perform semi-supervised clustering, cluster labels are mapped to class labels only at the evaluation stage, using the majority class in each cluster (2.32); the metrics are Accuracy, Precision, Recall, F1-Score, Specificity, RI, ARI and NMI.

$$\pi(G_j) = \arg\max_{c \in \mathcal{C}} \sum_{x_i \in G_j} \mathbb{I}(y_i = c) \tag{2.32}$$

2.5.2. Results on benchmark datasets

The parameter δ^{max} has a clear effect on clustering quality: small or moderately small δ^{max} values usually give higher and more stable Accuracy, whereas larger values let hyperboxes expand excessively, blurring cluster boundaries and reducing accuracy. The number of FMN-SAL hyperboxes decreases as δ^{max} increases; for example, on Aggregation it falls from 423.1 ($\delta^{max} = 0.01$) to 13.3 ($\delta^{max} = 0.80$). Table 2.2 shows that on the datasets for which both models report results, FMN-SAL achieves higher Accuracy than FMN-Voting; "-" denotes cases where FMN-Voting varied strongly with the order of training samples and Accuracy fell below 0.5, so no representative value is reported.

Table 2.2: *Accuracy of FMN-Voting and FMN-SAL*

Dataset	FMN-Voting	FMN-SAL
Flame	0.9622	0.9798
Iris	0.9588	0.9685
R15	0.9780	0.9984
Wine	0.9356	0.9450
Jain	-	0.9879
Spiral	-	0.9943
Pathbased	-	0.9781
PID	-	0.7004
Thyroid	-	0.9675
Aggregation	-	0.9911

Compared with the methods reported in [101] in terms of ARI (Table 2.4 of the dissertation), FMN-SAL obtains 0.9984 on R15 and 0.9798 on Flame, the highest values in the comparison; on Aggregation and Jain, FMN-SAL (0.9911 and 0.9879) is close to GCAE (0.9920 and 1.0000). Since the results of the compared methods are taken from [101] and were not re-run under the same protocol, they serve mainly as a reference.

Table 2.5 compares the Accuracy (%) of SCFMN, MSCFMN and FMN-SAL. FMN-SAL is higher on most datasets, notably 99.43% on Spiral. On PID and Sonar, SCFMN reaches 70.57% and 73.91%, slightly higher than 70.04% and 73.58% of FMN-SAL, showing that the benefit of sample-based voting depends on the characteristics of the data.

Table 2.5: *Accuracy (%) of three models on the datasets*

Dataset	δ^{max}	SCFMN	MSCFMN	FMN-SAL
Iris	0.06	94.12	94.67	96.85
Thyroid	0.07	92.65	92.63	96.43
Wine	0.06	93.33	93.86	96.78
PID	0.06	70.57	70.12	70.04
Sonar	0.06	73.91	73.54	73.58
Spiral	0.06	61.05	63.43	99.43
Flame	0.05	97.93	97.95	97.98
Jain	0.05	98.72	98.54	98.78
R15	0.05	99.50	99.64	99.84

2.5.3. Results on the ILPD liver disease dataset

On ILPD (liver disease as the positive class), both FMN-SAL and FMN-Voting perform best at $\delta^{max} = 0.08$. At this value, on the liver disease class FMN-SAL reaches Accuracy 0.9039, Recall 0.9423, Specificity 0.8084, Precision 0.9245 and F1-Score 0.9333, while FMN-Voting reaches Accuracy 0.8576, Recall 0.9135, Specificity 0.7186, Precision 0.8899 and F1-Score 0.9015. When δ^{max} is too large (0.90), Accuracy drops to about 0.51 and ARI and NMI are close to 0.

Tables 2.12 and 2.13 compare the proposed models with K-means++, EM, DBSCAN (run in Weka) and SCFMN. Precision, Recall and F1-Score are weighted averages over the two classes according to class sizes; for a single-label two-class problem, the weighted-average Recall equals Accuracy.

Table 2.12: *Overall evaluation on the ILPD dataset (extract)*

Model	Accuracy	Specificity	Precision	F1-Score
K-means++	0.6252 ± 0.0222	0.4463 ± 0.0932	0.6148 ± 0.0437	0.6154 ± 0.0240
EM	0.5647 ± 0.0351	0.7353 ± 0.1299	0.7589 ± 0.0544	0.5648 ± 0.0249
DBSCAN	0.6416 ± 0.0022	0.4411 ± 0.0048	0.6229 ± 0.0021	0.6306 ± 0.0019
SCFMN	0.7780 ± 0.0184	0.6442 ± 0.0330	0.7718 ± 0.0185	0.7736 ± 0.0186
FMN-SAL	0.8806 ± 0.0043	0.8001 ± 0.0082	0.8786 ± 0.0044	0.8788 ± 0.0043
FMN-Voting	0.8045 ± 0.0366	0.7029 ± 0.0490	0.8021 ± 0.0365	0.8030 ± 0.0364

The weighted-average Recall coincides with Accuracy and is therefore not repeated.

Table 2.13: *Comparison of FMN-SAL and FMN-Voting with other methods on ILPD*

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
SCFMN	0.6546 ± 0.0208	0.2772 ± 0.0432	0.1570 ± 0.0334
FMN-SAL	0.7894 ± 0.0066	0.5606 ± 0.0135	0.4077 ± 0.0137
FMN-Voting	0.6870 ± 0.0458	0.3499 ± 0.0942	0.2189 ± 0.0793

FMN-SAL achieves Accuracy 0.8806 ± 0.0043 and the highest RI, ARI and NMI in the comparison, with small standard deviations; FMN-Voting outperforms SCFMN and the Weka methods on most indices. K-means++ and DBSCAN have RI of about 0.53 but very low ARI and NMI,

indicating that their cluster structure does not match the two classes (with and without liver disease) well.

2.6. Conclusion of Chapter 2

Chapter 2 proposed two models, FMN-Voting and FMN-SAL. FMN-Voting exploits voting among neighbouring hyperboxes, reducing dependence on a single hyperbox; FMN-SAL infers labels from neighbouring labelled samples, reducing dependence on the geometric centre and representativeness of hyperboxes. The learning algorithms were presented with complexity analysis. Experiments show that both models achieve competitive results and improve performance on many datasets, although the degree of improvement depends on the dataset; the results on PID and Sonar indicate that no single label inference mechanism dominates on all datasets. The results of this chapter were published in [CT2], [CT3], [CT4] and [CT5].

CHAPTER 3. SEMI-SUPERVISED LEARNING IN FMN WITH AN ACTIVE SEED SELECTION STRATEGY

3.1. Introduction to the FA-Seed model

The models in Chapter 2 still depend on the quality of the initial seeds. Here, a seed is a representative sample selected and labelled to provide initial supervision; the reliability of a hyperbox, assessed through data density, the deviation between its geometric and data centres and the membership of its samples, reflects how well it can serve as a source for seed selection.

Existing semi-supervised FMN models either select seeds randomly (GFMM, SSL-FMM, SS-FMM) or select seeds from hyperboxes under the assumption that "each hyperbox represents one cluster" (SCFMN, MSCFMN). This assumption only holds for nearly spherical clusters of similar size; for datasets with complex structures such as Aggregation, Flame and Pathbased, a hyperbox must expand to cover the whole cluster, overlapping several clusters and losing purity. The dissertation therefore proposes the FA-Seed model [CT6], which combines active seed selection with the adaptive partitioning ability of FMN: a cluster is represented by

several hyperboxes, seed reliability is assessed and labels are propagated in a controlled manner.

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

3.2.1. Problem description

Let X be a dataset of M samples, each x_r mapped into C clusters, and let the hyperbox set be $B = L \uplus U$, consisting of labelled hyperboxes L and unlabelled hyperboxes U. FA-Seed addresses semi-supervised clustering when the training data are almost entirely unlabelled, with three goals: (i) selecting a high-quality and representative seed set; (ii) reducing the number of expert queries; and (iii) controlled label propagation to improve clustering quality. From the unlabelled data, FA-Seed determines a small but representative target seed set through candidate evaluation and active expert querying, then expands it in a controlled way into an expanded seed set that guides semi-supervised training.

3.2.2. Idea and architecture of FA-Seed

In the seed identification phase, FA-Seed uses a smaller hyperbox size limit to partition the data space into finer local regions (e.g. 41 hyperboxes on Aggregation, 11 on Flame and 19 on Pathbased). This constraint highlights local density variations and helps identify boundary regions, overlapping regions and outliers. The model has two phases (Figure 3.7): (i) active seed set construction - unsupervised FMN training, evaluation of candidate hyperboxes by data density and distribution, expert queries to confirm labels and controlled expansion; (ii) guided semi-supervised training with the expanded seed set. In the Aggregation example, 15 of the 25 hyperboxes are selected to generate seeds.

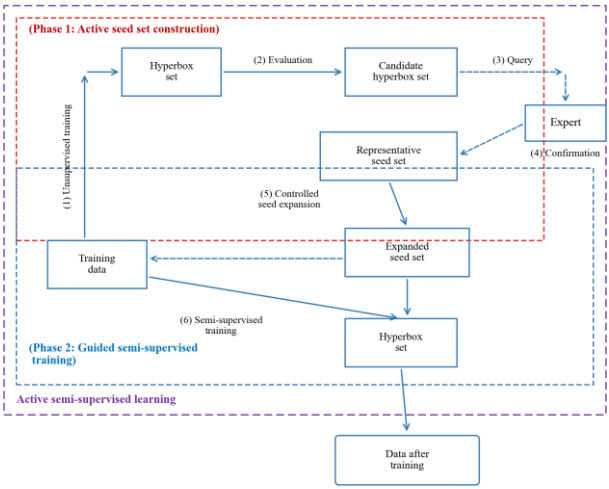


Figure 3.7: *Architecture of the fuzzy semi-supervised clustering model with active learning FA-Seed*

The FA-Seed network consists of the input layer F_A; the hyperbox evaluation layer F_B with K nodes corresponding to the K hyperboxes created in Phase 1; the label evaluation and propagation layer F_H, split into labelled and unlabelled hyperboxes, which applies gating mechanisms and label inheritance rules to move hyperboxes gradually into the labelled set; and the output layer F_C of cluster nodes.

3.2.3. Learning algorithm of FA-Seed

Seed set construction with active learning: the model actively selects representative samples from hyperboxes that meet the selection criteria and asks the expert to label them. Hyperboxes are evaluated by the centre deviation (2.1); impure hyperboxes are removed; the seed candidate set and the seed set are defined by

$$S = \bigcup_{j \in J_Q} \{x_r \in X \mid \mu(x_r, B_j) \geq \tau\},$$

$$J_Q = \{j \mid B_j \in \mathcal{B},\ e(B_j) \leq \varepsilon\}, \qquad Q = \{d(B_j) \mid j \in J_Q\} \qquad (3.9)\text{-}(3.12)$$

Algorithm 3.1 Seed set construction with active learning

Input: X ($M = |X|$); $\delta^{\max}$; ε

Output: Reordered dataset X', hyperbox set $\mathcal{B}$

Step 1: Generate the hyperbox set $\mathcal{B}$
Scan X once to build the fuzzy hyperbox set $\mathcal{B}$ based on $\delta^{\max}$;
Step 2: Filter hyperboxes (identify data points for querying)
for each hyperbox $B_j \in \mathcal{B}$ **do**
  Let T_j be the set of samples assigned to B_j (3.10); V_j, W_j are the min and max vertices of B_j;
  Compute the data centre $d(B_j)$ by (2.2) and the geometric centre $g(B_j)$ by (1.40);
  if condition (2.3) is not satisfied **then**
    Remove B_j: $\mathcal{B} \leftarrow \mathcal{B} \setminus \{B_j\}$;
  end if
end for
Step 3: Construct the seed set, partition and reorder the data
Construct the seed set Q from the data centres of the pure hyperboxes by (3.12) and query their labels from the expert;
Construct the seed set S from the samples lying entirely inside at least one pure hyperbox by (3.9);
Build a permutation π of $\{1, \ldots, M\}$ that places the elements of S first and $S' = X \setminus S$ afterwards;
Generate the reordered dataset $X' = (X_{\pi(1)}, \ldots, X_{\pi(M)})$;
return $\mathcal{B}$ (remaining hyperboxes), X' (reordered dataset)

Samples belonging to removed hyperboxes are not covered by any pure hyperbox, so they belong to *S′* and are processed in the next phase. Algorithm 3.1 has complexity $T_1 = O(nMK)$.

Guided semi-supervised training: the labelled seed set is used as initial supervision to guide hyperbox formation and label propagation; propagation is controlled by the hyperbox size constraint and the membership threshold β:

$$j^* = \arg\max_{j \in \mathcal{J}} \mu(x_r, B_j), \qquad \mathcal{J} = \left\{ j \mid \delta(x_r, B_j) \le \delta_2^{\max} \right\} \tag{3.19)-(3.20}$$

$$k^* = \arg\max_{k \in J_L} \mu\big(g(B_j), B_k\big), \qquad J_L = \left\{ k \mid \mu\big(g(B_j), B_k\big) \ge \beta \right\} \tag{3.22)-(3.23}$$

Algorithm 3.2 has complexity $T_2 = O(nMK)$. FA-Seed also incurs a query cost $O(q)$ for the q selected samples; since q is small, the overall order is unchanged. The total cost of FA-Seed is

$$T = T_1 + T_2 = O(nMK) + O(nMK) = O(nMK) \tag{3.27}$$

Algorithm 3.2 Guided semi-supervised training

Input: X'; β; $\delta_2^{\max}$
Output: $\mathcal{B}$

$\mathcal{L} = \varnothing; \mathcal{U} = \varnothing$;
for each $x_r \in X'$ **do**
  Determine the set $\mathcal{J}$ by (3.20);
  if $\mathcal{J} \neq \varnothing$ **then**
    Determine j^* by (3.19); expand B_{j^*} by (3.21);
    Adjust B_{j^*} and B_k by (1.15)–(1.26) if they overlap with $\ell(B_{j^*}) \neq \ell(B_k)$;
  else
    Create B_{new} with $V_{\text{new}} = W_{\text{new}} = x_r$;
    if $\ell_{x_r} \neq \varnothing$ **then** $\ell(B_{\text{new}}) = \ell_{x_r}$; $\mathcal{L} = \mathcal{L} \cup \{B_{\text{new}}\}$ **else** $\mathcal{U} = \mathcal{U} \cup \{B_{\text{new}}\}$;
  end if
end for
repeat
  for each $B_j \in \mathcal{U}$ **do**
    Compute the centroid $g(B_j)$ by (1.40);
    if $\exists k^*$ with $B_{k^*} \in \mathcal{L}$ satisfying (3.22) **then**
      $\ell(B_j) = \ell(B_{k^*})$; $\mathcal{L} = \mathcal{L} \cup \{B_j\}$; $\mathcal{U} = \mathcal{U} \setminus \{B_j\}$;
    end if
  end for
  Decrease β;
until ($\mathcal{U} = \varnothing$) **or** ($\beta \leq$ confidence threshold)
if $\mathcal{U} \neq \varnothing$ and $\mathcal{L} \neq \varnothing$ **then**
  For each $B_j \in \mathcal{U}$: $k^* \leftarrow \arg\max_k \mu(g(B_j), B_k)$, $B_k \in \mathcal{L}$; $\ell(B_j) = \ell(B_{k^*})$;
end if
$\mathcal{B} = \mathcal{L} \cup \mathcal{U}$
return $\mathcal{B}$

3.3. Experiments and evaluation

The experiments follow the setup of Chapter 2, with the Ecoli, Soybean and Zoo datasets added. On Iris and Wine at low label ratios (below about 20%), FA-Seed obtains a higher mean RI and smaller fluctuation than FC2; as the label ratio increases, the two methods converge. For the same number of queries, FA-Seed obtains more seeds than MSCFMN. In terms of RI, FA-Seed is generally better than or comparable to SSGC, SSDBSCAN, SSK-Means, K-Means and MSCFMN and reaches almost perfect RI on Soybean, Zoo and Iris; in terms of Accuracy, it is comparable to or higher than MSCFMN on most datasets, especially on imbalanced ones such as Thyroid, Pathbased and Zoo.

On ILPD, FA-Seed performs best at $\delta^{\max} = 0.08$, with Accuracy 0.9211, Recall 0.9567 and F1-Score 0.9454 on the liver disease class; RI,

ARI and NMI are 0.8544 ± 0.0263, 0.6972 ± 0.0549 and 0.5516 ± 0.0577, respectively. Compared with MSCFMN and the Weka methods (Tables 3.6 and 3.7), FA-Seed improves Accuracy from 0.7976 ± 0.0332 to 0.9009 ± 0.0165 and F1-Score from 0.7941 ± 0.0323 to 0.8998 ± 0.0167, while the FP Rate decreases from 0.3250 ± 0.0383 to 0.1638 ± 0.0255.

Table 3.6: *Comparison of FA-Seed with MSCFMN and Weka methods (extract)*

Model	Accuracy	Specificity	Precision	F1-Score
K-means++	0.6252 ± 0.0222	0.4463 ± 0.0932	0.6148 ± 0.0437	0.6154 ± 0.0240
EM	0.5647 ± 0.0351	0.7353 ± 0.1299	0.7589 ± 0.0544	0.5648 ± 0.0249
DBSCAN	0.6416 ± 0.0022	0.4411 ± 0.0048	0.6229 ± 0.0021	0.6306 ± 0.0019
MSCFMN	0.7976 ± 0.0332	0.6750 ± 0.0383	0.7929 ± 0.0329	0.7941 ± 0.0323
FA-Seed	0.9009 ± 0.0165	0.8362 ± 0.0255	0.8995 ± 0.0169	0.8998 ± 0.0167

The weighted-average Recall coincides with Accuracy and is therefore not repeated.

Table 3.7: *Comparison of FA-Seed with MSCFMN and Weka methods: RI, ARI, NMI*

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

Regarding computational cost, FA-Seed has complexity $O(nMK)$; when K is much smaller than M, the cost grows almost linearly with the number of samples, whereas SSGC has complexity $O(n^2)$. The current complexity analysis and running times only indicate potential computational efficiency and are not sufficient to claim scalability to large-scale data.

3.4. Conclusion of Chapter 3

Chapter 3 presented the FA-Seed model with an active seed selection strategy that jointly considers the representativeness and reliability of each seed, reducing the risk of label error propagation in overlapping, noisy or unevenly distributed regions at an acceptable computational cost. The experiments show that this strategy has the potential to improve clustering

quality. The results of this chapter were published in [CT6], [CT7] and [CT8].

CONCLUSION

1. Achievements

The dissertation studies semi-supervised clustering based on fuzzy min-max neural networks under limited label information, focusing on two issues: the reliability of label inference and propagation, and the quality of the initial seed set.

2. New contributions

(1) The FMN-Voting model, in which the labels of new samples and hyperboxes are inferred by majority voting among $k = 2C + 1$ neighbouring labelled hyperboxes instead of a single guiding hyperbox; a two-phase learning algorithm with complexity $O(R_{\max}nMK)$. Published in [CT3].

(2) The FMN-SAL model, which moves the voting source from hyperboxes to neighbouring labelled samples to reduce the influence of hyperboxes with large centre deviation or low usage index. On 7 of 9 benchmark datasets, FMN-SAL achieves higher Accuracy than SCFMN and MSCFMN (Table 2.5); on ILPD it reaches Accuracy 0.8806 ± 0.0043 and the highest RI, ARI and NMI in the comparison. Published in [CT2], [CT4] and [CT5].

(3) The FA-Seed model with active seed selection based on hyperbox partitioning with a small size limit, hyperbox evaluation by centre deviation and expert queries for representative samples, followed by guided semi-supervised training. On ILPD, FA-Seed reaches Accuracy 0.9009 ± 0.0165 and RI 0.8215 ± 0.0263, compared with 0.7976 ± 0.0332 and 0.6783 ± 0.0402 for MSCFMN. Published in [CT6], [CT7] and [CT8].

3. Scientific and practical significance

The dissertation incorporates structural information of hyperboxes (centre deviation, usage index, membership) into label inference and seed selection, with complexity analysis; the models are validated on the ILPD liver disease data, and the hyperbox structure allows interpretable IF-THEN rules to be extracted.

4. Limitations and future work

The dissertation has some limitations: (i) differences are reported as means and standard deviations without statistical significance tests, and on PID and Sonar FMN-SAL does not outperform SCFMN; (ii) FMN-Voting is not stable on some datasets with non-spherical clusters because it depends on the order of training samples; (iii) the performance depends on the hyperbox size limit δ^{max}; (iv) experiments are limited to tabular data with at most 7200 samples, without comparison with deep learning-based semi-supervised clustering; (v) FA-Seed requires expert participation in the query phase.

Future work includes: (i) combining the models with deep learning or feature representation to handle complex data; (ii) developing more adaptive label inference based on the reliability and dynamics of hyperboxes; (iii) parallel and distributed strategies to improve scalability; (iv) further exploiting fuzzy rule extraction to enhance interpretability in decision support systems; and (v) statistical significance testing and evaluation on larger datasets.

LIST OF PUBLICATIONS RELATED TO THE DISSERTATION

[CT1] Nguyen, T. S., Vu, T. D., Nguyen, A. T., Hoang, Q. H., & Vu, D. M. (2022). A Survey of Fuzzy Min-Max Neural Networks and Their Applications. In Selected Issues in Information and Communication Technology: Proceedings of the 25th National Conference, Hanoi, Vietnam, 8-9 December 2022 (pp. 238-244). Science and Technics Publishing House.

[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 the 2nd International Conference ICTA 2023, Volume 1, Lecture Notes in Networks and Systems, vol. 847, Springer Cham, 2023, pp. 134-141. DOI: 10.1007/978-3-031-49529-8_15.

[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", 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, 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, Can Tho City, Vietnam, October 18-19, 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, T. S., Vu, D. M., Le, B. D., & Nguyen, V. T. (2021). Semi-supervised clustering with a hyperbox weight-based seed selection method in fuzzy min-max neural networks. In Selected Issues in Information and Communication Technology: Proceedings of the 24th National Conference, Thai Nguyen, Vietnam, 13-14 December 2021 (pp. 433-438). Science and Technics Publishing House.

[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.