MINISTRY OF EDUCATION
AND TRAINING

VIETNAM ACADEMY OF
SIENCE AND TECHNOLOGY

GRADUATE UNIVERSITY OF SIENCE AND TECHNOLOGY

DUONG TIEN DUNG

RESEARCH ON A SEMI-SUPERVISED FUZZY CLUSTERING METHOD BASED ON BOUNDARY ANALYSIS AND ACTIVE LEARNING WITH PAIRWISE CONSTRAINTS

SUMMARY OF DISSERTATION ON COMPUTER
Code: 9 48 01 04

Ha noi - 2026

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

Supervisors:

Supervisor 1: Assoc. Prof. Dr. Ha Hai Nam.

Supervisor 2: Assoc. Prof. Dr. Nguyen Long Giang

Referee 1: …

Referee 2: …

Referee 3: ….

The dissertation will be examined by Examination Board of Graduate University of Science and Technology, Vietnam Academy of Science and Technology at…….. (time, date, year…)

This dissertation can be found at:

1) Graduate University of Science and Technology Library
2) National Library of Vietnam

Introduction

Motivation and significance of the research. Data clustering is a fundamental technique for knowledge discovery in large-scale and heterogeneous data (e.g., web, text, images), supporting latent structure modeling and dimensionality reduction [1]. Crisp clustering methods are limited when handling overlapping data, while Fuzzy C-Means (FCM), based on fuzzy set theory [2], assigns soft memberships and has been widely studied [3, 4]. However, FCM remains sensitive to initialization, noise, and errors near cluster boundaries.

Semi-supervised fuzzy clustering (SSFC) incorporates external knowledge (e.g., labels, pairwise constraints) to improve performance [5, 6], but still suffers from unreliable supervision, especially in ambiguous boundary regions. Active Learning addresses this by selectively querying informative samples, particularly boundary points, reducing labeling cost and improving reliability [7, 8]. This is especially relevant for applications such as image segmentation (e.g., medical imaging, remote sensing) [9, 10].

Motivated by these limitations, this dissertation integrates Active Learning into SSFC to enhance clustering performance under noisy, overlapping, and complex data conditions.

Research objectives. (1) To analyze and evaluate SSFC and Active Learning methods on benchmark datasets. (2) To propose Active Learning-based SSFC methods focusing on boundary regions and unreliable labels. (3) To apply the proposed methods to image segmentation tasks, particularly flood detection.

Research objects, scope, and methodology. The study focuses on SSFC and Active Learning-based fuzzy clustering under noisy data and ambiguous boundaries, evaluated on benchmark (e.g., UCI) and image datasets. Methods are assessed using standard clustering indices. The methodology includes: literature review, algorithm improvement, and experimental validation.

Main contributions. (1) Proposes an *Active Learning-based SSFC method guided by cluster boundaries* to refine ambiguous regions and improve clustering accuracy. (2) Proposes a *safe Active Learning-based SSFC method* with boundary-guided pairwise constraints to balance accuracy and error control.

These contributions enhance robustness and accuracy while reducing labeling costs.

Dissertation structure. *Chapter 1* reviews clustering, SSFC, Active Learning, and evaluation metrics, and identifies research gaps. *Chapter 2* presents the boundary-guided Active Learning SSFC method and experiments on benchmark and image data. *Chapter 3* develops a safe variant with pairwise constraints and applies it to flood detection. The conclusion summarizes findings and future directions.

Chapter 1

An Overview of Semi-supervised Fuzzy Clustering, Active Learning, and Practical Applications of Semi-supervised Clustering

1.1 Overview of Semi-supervised Fuzzy Clustering

1.1.1 Fuzzy clustering

Fuzzy clustering (Fuzzy/Soft Clustering) extends "hard" clustering by assigning each data point a *membership vector* over c clusters, whose components lie in $[0, 1]$ and are typically normalized to sum to 1. This soft representation reflects practical data characteristics more faithfully when *cluster boundaries are ambiguous*, *clusters overlap*, or the data are *noisy*, since a sample may simultaneously belong to multiple clusters to different degrees.

1.1.2 Semi-supervised clustering

Semi-supervised clustering (SSC) combines large-scale unlabeled data with a limited amount of *supervisory information* in order to improve clustering quality, reduce labeling cost, and enhance model stability compared with purely unsupervised learning. Common forms of supervision include: *(i) sparse labels* for a subset of samples [9]; *(ii) pairwise constraints* of the *must-link/cannot-link* type, specifying whether two samples should or should not belong to the same cluster [11]; and *(iii) predefined memberships* within the fuzzy clustering framework [12, 13].

1.1.3 Semi-supervised fuzzy clustering

In this review, FCM is regarded as the foundation of fuzzy clustering: the algorithm alternately optimizes sample memberships and cluster centers, and has demonstrated effectiveness across a wide range of domains, including image segmentation, pattern recognition, and medical analysis. Nevertheless, it remains sensitive to initialization, vulnerable to noise and outliers, and implicitly assumes nearly isotropic cluster geometry [14]. Numerous variants have therefore been developed to improve robustness and flexibility, including rough/kernel/robust formulations, Wasserstein-based methods, spatial weighting schemes, fuzzifier selection strategies, hybrid optimization approaches, eigenvalue-constrained models, and anti-monotonic regularization terms [15].

When labels are scarce but semantic consistency is required, semi-supervised fuzzy clustering (SSFC) incorporates domain knowledge into the objective function through sparse labels, ML/CL constraints, and predefined memberships. Kernel-based SSFC

methods, including multi-kernel variants, improve the modeling of nonlinear structures but are often difficult to tune; adaptive branches are designed for streaming and high-dimensional data; and "safe" variants mitigate the influence of unreliable labels through confidence weighting and local graph structures [16–18]. Notably, TS3FCM filters reliable labels, initializes memberships, and then performs global optimization, thereby improving performance; however, it remains affected by initialization and noisy boundary regions [12].

In the direction of deep integration, neuro-fuzzy and Deep SSFC models jointly learn data representations (e.g., by AE/CNN) together with FCM-type constrained objectives, achieving promising performance on complex data, though at the expense of hyperparameter sensitivity and substantial training costs [19, 20]. Finally, Active Learning for SSFC selects queries or pairwise constraints from uncertain boundary regions—for example, AFCC (FHV), K-GBS^{3}FCM ("safe" KNN graph), and SFCM-PM (prior + entropy)—thus enabling high clustering quality with very limited supervision, albeit still subject to parameter sensitivity and graph-construction costs [5, 8, 21]. These observations indicate several unresolved issues: the tight integration of active pair selection with safe mechanisms against label noise, the joint optimization of representation–constraint–sample selection, and the design of methods with lower parameter sensitivity and better scalability to large datasets.

1.2 Active Learning

Active Learning aims to reduce labeling cost by *actively* selecting the most informative samples for expert annotation, rather than collecting or labeling data randomly as in *passive learning*. Its central assumption is that, if the learner is allowed to choose which samples to query, it can achieve comparable or even superior performance with *fewer labeled instances*. This paradigm is particularly valuable when raw data are abundant but labels are scarce or expensive, and has proven effective in text classification, information extraction, image analysis, remote sensing, biomedicine, sensor selection, network structure learning, and software analysis [22].

Query types and selection criteria. Three common types of queries are typically considered: (i) *point-label queries*; (ii) *pairwise constraints* of the ML/CL type, which are widely used because they are intuitive for human annotators [7]; and (iii) *structural/comparative queries* such as triplets or "peak" queries in density peak clustering [23, 24]. Representative utility criteria include: *uncertainty* (e.g., entropy, margin, committee disagreement), *representativeness* (e.g., density, kNN structure), *influence/expected model change*, and *diversity* to avoid redundancy in batch querying; in practice, combining multiple criteria is often more effective than relying on any single one [25–27].

1.3 Recent studies on semi-supervised fuzzy clustering

1.3.1 Active semi-supervised fuzzy clustering

A. Context and motivation

Semi-supervised fuzzy clustering (SSFC) combines unlabeled data with auxiliary information such as point labels, ML/CL constraints, or predefined memberships, thereby improving clustering accuracy compared with purely unsupervised fuzzy clustering [7, 27, 28]. However, SSFC still faces several challenges, including high labeling cost, noise and boundary points that distort cluster structure, sensitivity to parameters and the number of clusters, the possibility of contradictory constraints, and substantial computational cost. Active Learning is regarded as a promising solution because it queries "few but informative" samples, focusing on uncertain regions in order to reinforce cluster boundaries and improve model robustness [25, 26].

B. Query strategies

In SSFC, supervision may take the form of direct labels (when the number of clusters is known), ML/CL constraints that are user-friendly [7, 8], or comparative structural information such as triplets/peaks suitable for density-based models [24]. Query selection is primarily based on four criteria: point uncertainty, representativeness, diversity, and model impact [25].

C. Representative studies

Important research directions include the following. AFCC (2008) employs Fuzzy Hypervolume to select constraints in boundary regions, achieving high accuracy but with substantial computational cost and parameter sensitivity [8]. ASC (2017) optimizes constraints on a KNN graph to estimate the number of clusters and reduce redundant queries [29]. K-GBS3FCM (2024) performs safe label propagation based on KNN structures, but depends strongly on the choice of K [21]. SFCM-PM (2025) incorporates predefined memberships, which is beneficial for imbalanced data but may become noise-sensitive when the prior information is unreliable [5].

D. Common limitations and research gaps

Although these methods reduce query cost, most still fail to handle highly overlapped boundary regions effectively—the areas where clustering errors are concentrated and where fuzzy models exhibit the highest uncertainty [27]. In addition, the optimization of queries based on membership structure remains limited [8], and very few methods place substantial emphasis on robustness to noise and outliers [25]. This constitutes a significant research gap.

E. Direction and contributions of the dissertation

This dissertation aims to exploit the fuzzy structure of clusters by explicitly identifying boundary regions through uncertainty and density information, and then designing an active querying strategy that prioritizes samples located near cluster boundaries—where expert feedback is most valuable. On this basis, the dissertation proposes an active semi-supervised fuzzy clustering algorithm guided by boundary regions, with the goals of reducing labeling cost, improving clustering accuracy, and enhancing ro-

bustness to noise, thereby providing greater effectiveness in practical computer vision applications.

1.3.2 Safe semi-supervised fuzzy clustering

A. Context and motivation

In semi-supervised fuzzy clustering (SSFC), the reliability of labeled data and ML/CL constraints plays a decisive role. When these sources of information are noisy, inconsistent, or biased, the model may easily amplify errors through label propagation, be misdirected during optimization by assigning excessive weight to unreliable samples, or become trapped in local optima near cluster boundaries. For this reason, safe learning in SSFC emphasizes the identification and regulation of reliability at the sample, pair, and cluster levels, in order to maintain stability throughout the optimization process.

B. Recent safe SSFC approaches

Recent studies have incorporated weighting and adaptive mechanisms to improve model stability. S^3FCM, LHC-S3FCM, and CS3FCM [18, 30] employ confidence measures combined with neighborhood graphs to adjust the contribution of labeled samples. TS3FCM [12] further improves convergence and robustness by filtering labels and exploiting predefined memberships, and is particularly effective in noisy data environments.

C. Characteristics and limitations

Although these methods significantly improve safety, they still depend strongly on random initialization and nonuniform density distributions, remain ineffective in highly overlapped boundary regions or noisy samples, and provide only limited control over constraint-related risks when fuzzy clusters intersect.

D. Research gap

There is still a lack of mechanisms that guarantee *local* safety in ambiguous boundary regions, where errors are most heavily concentrated. Moreover, the reliability of labels and constraints has not yet been naturally integrated into the fuzzy structure itself, making it difficult for the model to balance safety and supervision cost—particularly in highly noisy problems.

E. Direction and contributions of the dissertation

This dissertation aims to identify and process ambiguous boundary regions based on uncertainty and density; to design a self-adjusting weighting mechanism according to reliability levels in order to limit the impact of noise; and to develop a safe and effective SSFC model for computer vision applications, where reliability and noise tolerance are especially critical. The ultimate objective is to improve robustness to noise and clustering quality in complex real-world scenarios.

1.4 Models and evaluation data

To verify effectiveness under conditions of fuzzy boundaries, severe overlap, and realistic noise, this dissertation employs three complementary data sources: (i) benchmark UCI datasets (10 classical datasets, including Iris, Wine, BCW, Glass, Thyroid,

Soybean(small), Haberman, Australian Credit, Spambase, and Waveform-5000) for quantitative comparison against standard baselines [31]; (ii) synthetically generated data with controllable overlap (GEN2D), designed to isolate and regulate the overlap level (ratio p) and boundary fuzziness (mixing coefficient α), thereby supporting sensitivity analysis with respect to noise, the number of clusters, and density bandwidth; and (iii) image datasets, including (a) Landsat-8 images from the coastal area of Thanh Hoa for land-cover classification (6 classes), characterized by complex boundaries and spectral heterogeneity; and (b) a flood-region segmentation dataset comprising approximately 290 images with binary masks ("flood"/"non-flood") for evaluation in scenarios involving thin boundaries and reflection noise.

Evaluation metrics. The dissertation employs the following set of measures: RI, F1, NMI, DB, PC, PE, and DB_fuzzy.

1.5 Conclusion of Chapter 1

Chapter 1 has systematized the theoretical foundations for image segmentation, beginning with fuzzy sets, the role of membership degrees, and the FCM algorithm, and extending to semi-supervised fuzzy clustering (SSFC) with seed labels, scribbles, ML/CL constraints, and neighborhood information. The chapter has clarified the role of Active Learning in SSFC—namely, the selection of point, pairwise, or structural queries from boundary regions, representative samples, or highly influential samples—with the aim of substantially reducing labeling cost. It has also established links with deep learning approaches (e.g., FCN, U–Net, DeepLab, etc.), which enable the preservation of complex boundaries while reducing annotation effort.

A unified set of evaluation measures is employed throughout the dissertation, including RI, pairwise F1, and NMI (external criteria), together with DB (an internal criterion), in order to assess both label agreement and the compactness–separation characteristics of clusters. For fuzzy data, RI, F1, and NMI are computed after defuzzification using the maximum-membership rule, whereas fuzzy clustering measures such as PC and PE are used to reflect cluster separability and the overall degree of fuzziness in the system. The DB_fuzzy index is employed alongside DB to characterize cluster compactness in the membership space, which is particularly useful when clusters are separated by soft boundaries.

From this review, several remaining research gaps can be identified, particularly those related to preserving boundary regions in the fuzzy space, improving label reliability, and developing large-scale active querying mechanisms. These issues provide the motivation for the methods and models presented in the subsequent chapters.

From the review, the unresolved issues of boundary preservation, label reliability, and large-scale active querying are identified as the main motivations for the following chapters.

Chapter 2

A Proposed Active Semi-supervised Clustering Method Based on Cluster Boundaries

2.1 Cluster boundaries and issues in semi-supervised fuzzy clustering

Cluster boundaries are regions in which data memberships are ambiguous and thus highly prone to mislabeling—particularly in image and biological data—making SSFC, even when equipped with pairwise constraints, still sensitive to initialization and difficult to apply effectively in the presence of overlap and noise. Active Learning can improve this process by prioritizing queries on the most "informative" samples located precisely in boundary regions and by applying ML/CL constraints to stabilize these boundaries at low labeling cost. This requires a reliable mechanism for boundary identification, robust query criteria under noisy conditions, and a fuzzy objective function with a *soft penalty* term to mitigate the effect of incorrect supervision, thereby leading to an SSFC+AL framework following a *boundary–centroid* strategy.

2.2 Main idea and algorithmic details

The proposed algorithm is built upon three main ideas, all centered on *handling boundary regions*, where data memberships are ambiguous: (i) an initial FCM run is performed to obtain a soft partition and to *detect boundary regions*—namely, points whose memberships are distributed across multiple clusters; (ii) **Active Learning** is then applied only to these boundary regions, where ambiguous samples are selectively queried from an oracle in order to refine labels/memberships, thereby stabilizing local boundaries at low labeling cost; and (iii) the clustering solution is subsequently *refined* by means of a new objective function that exploits the corrected boundary points and the resulting "boundary centroids."

The overall algorithm consists of five main steps. First, the model is initialized by running FCM over the entire dataset in order to obtain the initial membership matrix $U^{(0)}$ and the set of cluster centers $V^{(0)}$, which serve as the basis for the subsequent semi-supervised procedure. Next, boundary regions are identified by measuring the ambiguity of each sample, for example through the discrepancy index $\Delta_k = \min_{p \neq q} |u_{kp} - u_{kq}|$, and the N_q most uncertain samples are selected as boundary candidates. Then, an Active Learning mechanism is applied to these samples: the model queries an oracle to obtain labels or prior information, from which a predefined membership matrix $\bar{U}$ is constructed to encode reliable local supervisory knowledge. Based on the supervised boundary samples, the subsequent semi-supervised step optimizes the proposed

objective function by adjusting memberships so that they approach the preferred values in $\bar{U}$. Finally, the algorithm iteratively updates U and V until the cluster centers converge (with error not exceeding ε) or the maximum number of iterations is reached.

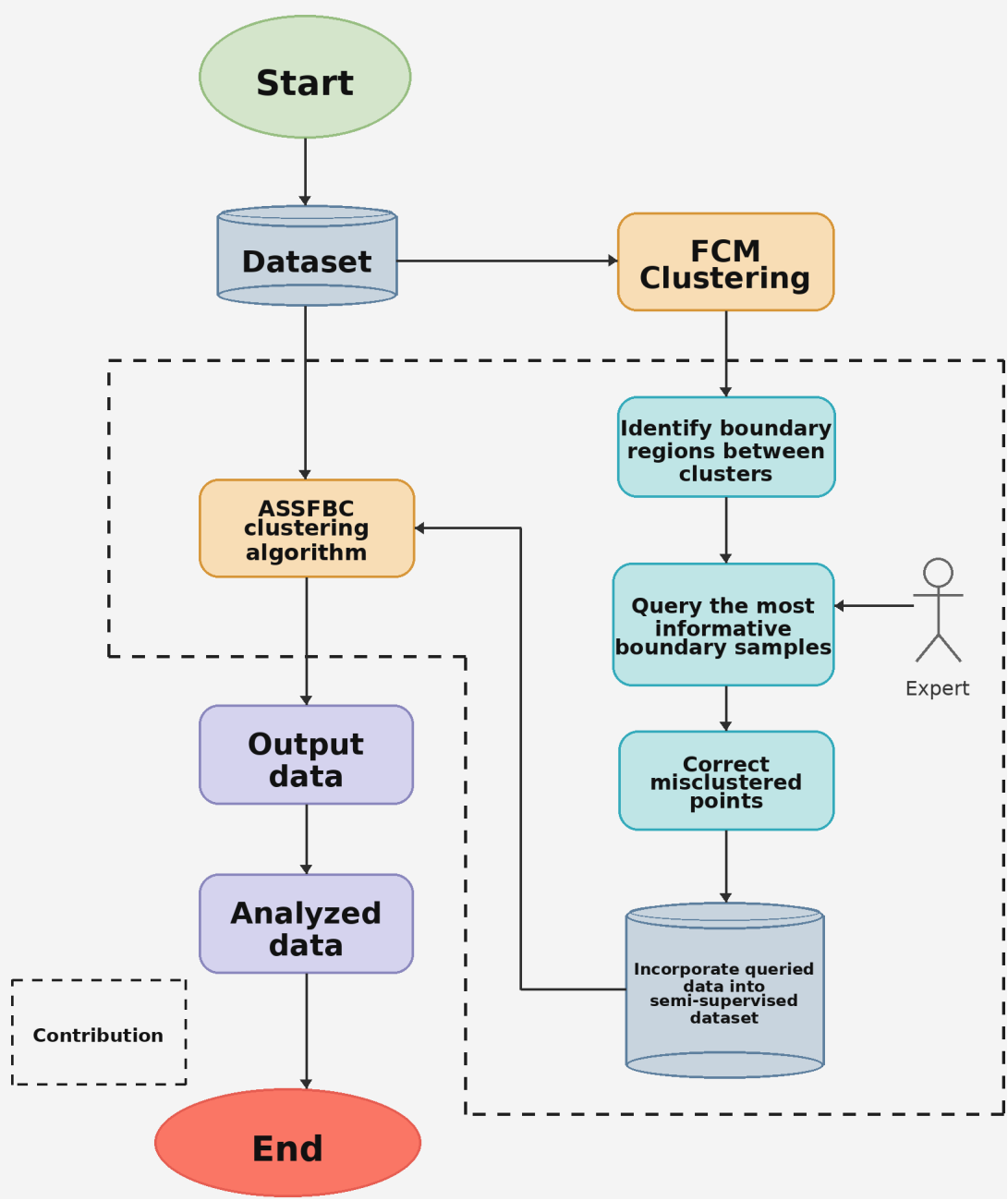

Figure 2.1: Active semi-supervised fuzzy clustering model based on cluster boundary regions (ASSFBC).

Objective function

$$\begin{aligned}\min_{u,v}\ J(u,v) = &\sum_{k=1}^{N}\sum_{j=1}^{C} u_{kj}^2 \left\|x_k - v_j\right\|^2 + \alpha \sum_{k=1}^{L}\sum_{j=1}^{C} \left|u_{kj} - \bar{u}_{kj}\right|^2 \left\|x_k - v_j\right\|^2 \\ &+\beta \sum_{i=1}^{C-1}\sum_{d=i+1}^{C}\sum_{l\in N_{id}} (1-\mu_{id}^{l}) \left(\left\| v_i - \tfrac{1}{|N_i|}\sum_{h\in N_i} x_h \right\|^2 + \left\| v_d - \tfrac{1}{|N_d|}\sum_{h\in N_d} x_h \right\|^2 \right)\end{aligned} \tag{2.1}$$

Parameter update

Membership. Subject to the constraint $\sum_j u_{kj} = 1$:

$$u_{kj} = \frac{\alpha \bar{u}_{kj}}{1+\alpha} + \frac{1 - \dfrac{\alpha}{1+\alpha} \sum\limits_{i=1}^{C} \bar{u}_{ki}}{d_{kj}^2 \sum\limits_{i=1}^{C} \frac{1}{d_{ki}^2}}. \tag{2.2}$$

Cluster centers.

$$v_j = \frac{2 \sum\limits_{k=1}^{N} u_{kj}^2 x_k + 2\alpha \sum\limits_{k=1}^{L} |u_{kj} - \bar{u}_{kj}|^2 x_k + 2\beta \sum\limits_{l \in N_{id}} (1 - \mu_{id}^l) T_{id}}{2 \sum\limits_{k=1}^{N} u_{kj}^2 + 2\alpha \sum\limits_{k=1}^{L} |u_{kj} - \bar{u}_{kj}|^2 + 2\beta \sum\limits_{l \in N_{id}} (1 - \mu_{id}^l)}. \tag{2.3}$$

Algorithm 1 ASSFBC

1: **Input**: Dataset X, number of samples N, number of clusters C, `seed_ratio`, convergence threshold ε, maximum number of iterations $T_{\max}$
2: (FCM) Initialize the membership matrix $U^{(0)}$ and cluster centers $V^{(0)}$ using the FCM algorithm.
3: (Boundary) Compute the ambiguity measure Δ_k and select $N_q = \lfloor \texttt{seed\_ratio} \cdot N \rfloor$ samples from the boundary regions.
4: (Active Learning) Query the oracle $\Rightarrow$ refine $\bar{U}$ on the boundary samples.
5: Set $t \leftarrow 1$.
6: **Repeat** while $t \le T_{\max}$:
7: Update $U^{(t)}$ according to Eq. 2.2.
8: Update $V^{(t)}$ according to the cluster-center update formula in Eq. 2.3.
9: **If** $\max_j \|v_j^{(t)} - v_j^{(t-1)}\| \le \varepsilon$ **then** stop.
10: Set $t \leftarrow t + 1$.
11: **End repeat**

Complexity

With I_0 initialization iterations for FCM, T optimization iterations, and $N_q \ll N$:

$$\boxed{\text{Time} \;=\; O\big(NC^2(I_0 + T) + O(N \log N)\big)}$$

2.3 Experimental results

To demonstrate the effectiveness of the proposed method in situations involving substantial noise in cluster boundary regions, experiments were conducted to compare the proposed approach under three categories of data: benchmark UCI datasets, manually generated datasets, and image data.

The experiments were implemented in MATLAB on an LG Gram laptop equipped with an Intel(R) Core(TM) i5-6200U CPU @ 2.30GHz to 2.40GHz and 8GB RAM. Comparative experiments were conducted between the proposed ASSFBC method and several related approaches, including: FCM [4], SSFCM [29], eSFCM [44], AFFC [45], and AFFC [46]. The clustering quality was evaluated using the following indices: RI, F1, NMI, DB, PC, PE, and DB_fuzzy. After conducting the experiments, the results

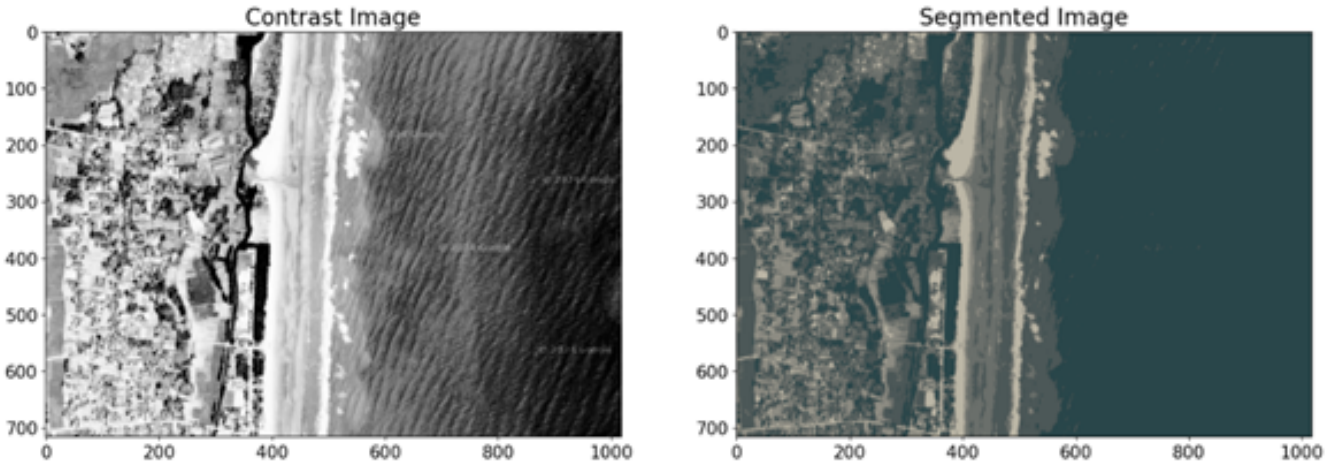

Figure 2.2: Illustration of land-cover segmentation on Landsat-8 imagery

are reported in detail in **Tables 2.1, 2.2, and 2.3**, corresponding respectively to the **UCI datasets, synthetic data, and image data**.

Algorithm Index	**FCM**	**SSFCM**	**eFCM**	**AFFC(2008)**	**AFFC(2017)**	**ASSFBC**
IRIS						
RI	0.8741	0.9018	0.9094	0.9253	0.9492	**0.9939**
F1	0.8364	0.8491	0.8614	0.9086	0.9304	**0.9898**
NMI	0.7381	0.7704	0.7825	0.8241	0.8463	**0.9744**
DB	**0.9414**	1.3615	1.4173	1.5471	1.6932	1.2294
PC	0.7803	0.8112	0.8324	0.8517	0.8831	**0.9418**
PE	0.4216	0.3897	0.3712	0.3431	0.3124	**0.2135**
DB_{fuzzy}	**0.3827**	0.4721	0.5834	0.5219	0.5028	0.4617
Breast						
RI	0.7594	0.7594	0.7641	0.7689	0.7731	**0.7848**
F1	0.7894	0.7894	0.7961	0.8044	0.7971	**0.8667**
NMI	0.4751	0.4751	0.4928	0.4974	0.5002	**0.5123**
DB	**0.7485**	0.8156	1.3481	2.3923	0.9658	0.9144
PC	0.7412	0.7427	0.7529	0.7638	0.7714	**0.7925**
PE	0.4829	0.4817	0.4698	0.4523	0.4427	**0.4235**
DB_{fuzzy}	0.7124	0.6942	**0.6028**	0.7729	0.8157	0.9526
Glass						
RI	0.8214	0.8192	0.8258	0.8310	0.8322	**0.8494**
F1	0.5983	0.5942	0.6029	0.6058	0.6091	**0.6154**
NMI	0.6891	0.6804	0.6875	0.6971	0.6582	**0.7584**
DB	**0.6383**	2.8294	1.9192	1.4598	2.1304	1.1964
PC	0.6921	0.6834	0.7019	0.7128	0.7256	**0.7517**
PE	0.5318	0.5524	0.5247	0.4928	0.4719	**0.4125**
DB_{fuzzy}	**0.5231**	1.8827	0.7216	1.0029	1.5324	2.2612

Table 2.1: Overall comparison of evaluation metrics on the IRIS, Breast, and Glass datasets.

Algorithm Index	FCM	SSFCM	eFCM	AFFC(2008)	AFFC(2017)	ASSFBC
Wine						
RI	0.7199	0.7418	0.7367	0.7556	0.7583	**0.7689**
F1	0.5868	0.6234	0.6166	0.6417	0.6459	**0.6578**
NMI	0.4295	0.4537	0.4469	0.4604	0.4671	**0.4955**
DB	**0.8365**	1.6125	0.8490	0.8367	0.9608	2.2867
PC	0.7213	0.7328	0.7394	0.7517	0.7596	**0.7729**
PE	0.5074	0.4931	0.4989	0.4728	0.4627	**0.4382**
DB_{fuzzy}	**0.6934**	1.5142	0.7249	0.8157	0.9345	2.1479
Soybean						
RI	0.8234	0.8408	0.7865	0.8441	0.8464	**0.8677**
F1	0.6819	0.7124	0.6081	0.7168	0.7194	**0.7571**
NMI	0.7002	0.7544	0.5454	0.7568	0.7671	**0.7784**
DB	3.7219	3.7587	4.7462	**2.4579**	2.7805	2.7581
PC	0.8124	0.8317	0.7829	0.8368	0.8441	**0.8619**
PE	0.3881	0.3627	0.4786	0.3579	0.3425	**0.3287**
DB_{fuzzy}	3.4972	3.6045	4.5289	**2.3841**	2.6974	2.6812
Thyroid						
RI	0.6394	0.6513	0.6655	0.7175	0.7258	**0.7649**
F1	0.6314	0.6338	0.6485	0.6872	0.6988	**0.7367**
NMI	0.3135	0.2968	0.3301	0.3635	0.3671	**0.5284**
DB	**2.0871**	2.4405	4.8259	2.6492	2.5697	2.1271
PC	0.6329	0.6463	0.6617	0.7018	0.7163	**0.7524**
PE	0.5871	0.5783	0.5692	0.5393	0.5284	**0.4189**
DB_{fuzzy}	**2.0471**	3.3829	4.6985	2.7987	2.5234	2.4481

Table 2.2: Overall comparison of evaluation metrics on the Wine, Soybean, and Thyroid datasets.

Based on the experimental results, ASSFBC consistently outperforms competing methods in terms of RI, F1, and NMI, as well as fuzzy measures such as PC and PE, across most datasets. These two fuzzy measures indicate that the membership matrix becomes sharper (higher PC) while the degree of fuzziness is substantially reduced (lower PE), reflecting the fact that correcting boundary regions helps reduce ambiguity in difficult-to-classify samples. For internal indices such as DB and DB_fuzzy, the values of ASSFBC are sometimes higher than those of FCM. This can be explained by the fact that FCM optimizes only the point-to-center distance, thereby naturally producing very compact, symmetric, and spherical clusters, which leads to lower DB and DB_fuzzy values. Nevertheless, compared with other semi-supervised fuzzy clustering algorithms, ASSFBC generally achieves better or competitive DB and DB_fuzzy values, indicating that the adjustment of boundary regions does not deteriorate the cluster structure and still ensures stability.

On IRIS, Breast, Glass, Wine, Soybean, Thyroid, Data2, and Satellite, ASSFBC achieves the best performance on all three label-based measures; in particular, on Thyroid it also yields the lowest DB value, showing that boundary correction can preserve cluster compactness. On several datasets such as Glass, Wine, Data2, Satellite, and Breast, higher DB values reflect the tendency of clusters to become slightly

more dispersed when the algorithm prioritizes boundary correction in order to better match labels under substantial overlap. However, this does not significantly affect the label-based clustering quality. Overall, ASSFBC is stable and particularly strong with respect to label-based measures, while its high PC, low PE, and competitive DB_fuzzy values indicate a favorable balance between membership sharpness and cluster compactness. Further reductions in DB and DB_fuzzy could potentially be achieved by incorporating an additional compactness term, learning an adaptive distance metric, or refining the pair-selection strategy.

Algorithm Index	**FCM**	**SSFCM**	**eFCM**	**AFFC(2008)**	**AFFC(2017)**	**ASSFBC**
Data1						
RI	0.9523	0.9725	0.8941	**1.0000**	**1.0000**	**1.0000**
F1	0.9320	0.9624	0.8927	**1.0000**	**1.0000**	**1.0000**
NMI	0.9013	0.9317	0.7094	**1.0000**	**1.0000**	**1.0000**
DB	1.8451	2.3049	3.0185	**0.9282**	0.9347	0.9433
PC	0.9321	0.9340	0.8875	0.9687	0.9714	**0.9751**
PE	0.1184	0.1127	0.1548	0.0417	0.0372	**0.0294**
DB_{fuzzy}	1.0500	1.2800	1.6500	**0.4800**	0.5100	0.5200
Data2						
RI	0.7194	0.7426	0.7892	0.8131	0.8378	**0.8725**
F1	0.6694	0.6825	0.7024	0.7529	0.7391	**0.8148**
NMI	0.4895	0.5731	0.6152	0.6418	0.6184	**0.6761**
DB	**0.7724**	1.0081	1.5849	3.6928	4.5824	4.3715
PC	0.7216	0.7429	0.7714	0.7948	0.8162	**0.8467**
PE	0.5174	0.4831	0.4518	0.4236	0.4379	**0.3927**
DB_{fuzzy}	**0.6800**	0.9100	1.4800	3.4500	4.3150	4.1800
Satellite Image						
RI	0.7068	0.7309	0.7162	0.8278	0.8367	**0.8589**
F1	0.6847	0.6968	0.7061	0.7953	0.8075	**0.8368**
NMI	0.6314	0.6456	0.6593	0.7289	0.7494	**0.7879**
DB	0.5978	0.6081	**0.5506**	0.7371	0.7268	0.7498
PC	0.7051	0.7283	0.7152	0.8257	0.8354	**0.8609**
PE	0.5258	0.5006	0.4909	0.3952	0.3805	**0.3457**
DB_{fuzzy}	0.6108	0.6204	**0.5653**	0.7457	0.7324	0.7608

Table 2.3: Overall comparison of evaluation metrics on the synthetic datasets and image data.

2.4 Conclusion

This dissertation has introduced a novel approach to active semi-supervised fuzzy clustering, with particular emphasis on refining cluster boundary regions. The proposed method effectively addresses the challenge of uncertainty in cluster boundaries by integrating Active Learning techniques to improve the accuracy of membership assignment at critical samples. This approach not only enhances the reliability of clustering results but also establishes a new semi-supervised fuzzy clustering framework that tightly integrates fuzzy clustering and Active Learning for boundary refinement.

Experimental results on benchmark datasets demonstrate that the proposed ASSFBC

method achieves superior performance compared with traditional fuzzy clustering algorithms and other semi-supervised clustering methods. This superiority is reflected in consistently higher RI, F1, and NMI values, indicating better label agreement and more accurate cluster separation. In addition, fuzzy measures such as PC and PE are substantially improved: the increase in PC shows that the membership matrix becomes sharper, while the reduction in PE indicates that fuzziness is narrowed in difficult-to-classify regions. For internal measures such as DB and DB_fuzzy, the values obtained by ASSFBC are generally lower than or competitive with those of most other semi-supervised methods, although not always lower than FCM. The reason is that FCM optimizes only the point-to-center distance, thereby naturally generating highly symmetric and compact clusters, which leads to lower DB and DB_fuzzy values. Nevertheless, ASSFBC still demonstrates a favorable balance between cluster compactness and label-based accuracy, highlighting the benefit of boundary adjustment in improving overall clustering quality.

The results presented in this chapter have been published in **CT1, CT2**.

Chapter 3

Proposed Active Safe Semi-Supervised Clustering Method with Pairwise Links Based on Cluster Boundaries

3.1 Algorithmic Idea

The model integrates *fuzzy clustering* (to represent uncertainty at cluster boundaries), *semi-supervised learning* (to exploit limited labels/constraints), and *active learning* (to select the most informative queries in ambiguous regions). The core idea is to use pairwise ML/CL constraints confirmed by an *oracle* at cluster boundaries in order to adjust membership values and update cluster centers, thereby narrowing overlap regions, increasing inter-cluster separability, and improving robustness against noisy labels. The iterative procedure is as follows: FCM initialization → boundary region detection → query/apply ML/CL constraints → re-optimize clustering. This integration achieves high accuracy when labels are scarce and the data structure is complex, while preserving interpretability through domain-informed constraints.

3.2 Methodology

The proposed **AS3FCPC** method consists of four interrelated stages designed to reduce uncertainty in *cluster boundary regions* under label scarcity: (i) *initialization by active learning + FCM on the labeled subset*, (ii) *iterative fuzzy semi-supervised clustering* on the whole dataset, (iii) *boundary refinement by active learning and selective generation of pairwise constraints*, (iv) *iterative optimization until convergence*.

The key principle is to *only* query and adjust ambiguous samples (whose membership values are very close), and then *propagate* this information through the semi-supervised objective and ML/CL constraints.

The algorithm is implemented through four consecutive stages. In the first stage, active learning is applied to the labeled subset by identifying the neighborhood of each sample through the radius $R = (d_{\max} - d_{\min})\,\alpha$. A sample is regarded as ambiguous when fewer than 50% of its neighbors share the same label; conversely, if the majority of its neighbors belong to another label, the sample label is corrected according to the majority rule. After this filtering step, FCM is applied to obtain stable initial U and V, which serve as the foundation for the semi-supervised stage.

In the second stage, fuzzy semi-supervised clustering is performed on the entire dataset by using $\bar{U}$ from the previous stage as supervisory anchors. The update rules for U and V follow the FCM principle combined with a penalty term that reduces the deviation from $\bar{U}$, thereby propagating reliable label information to unlabeled data.

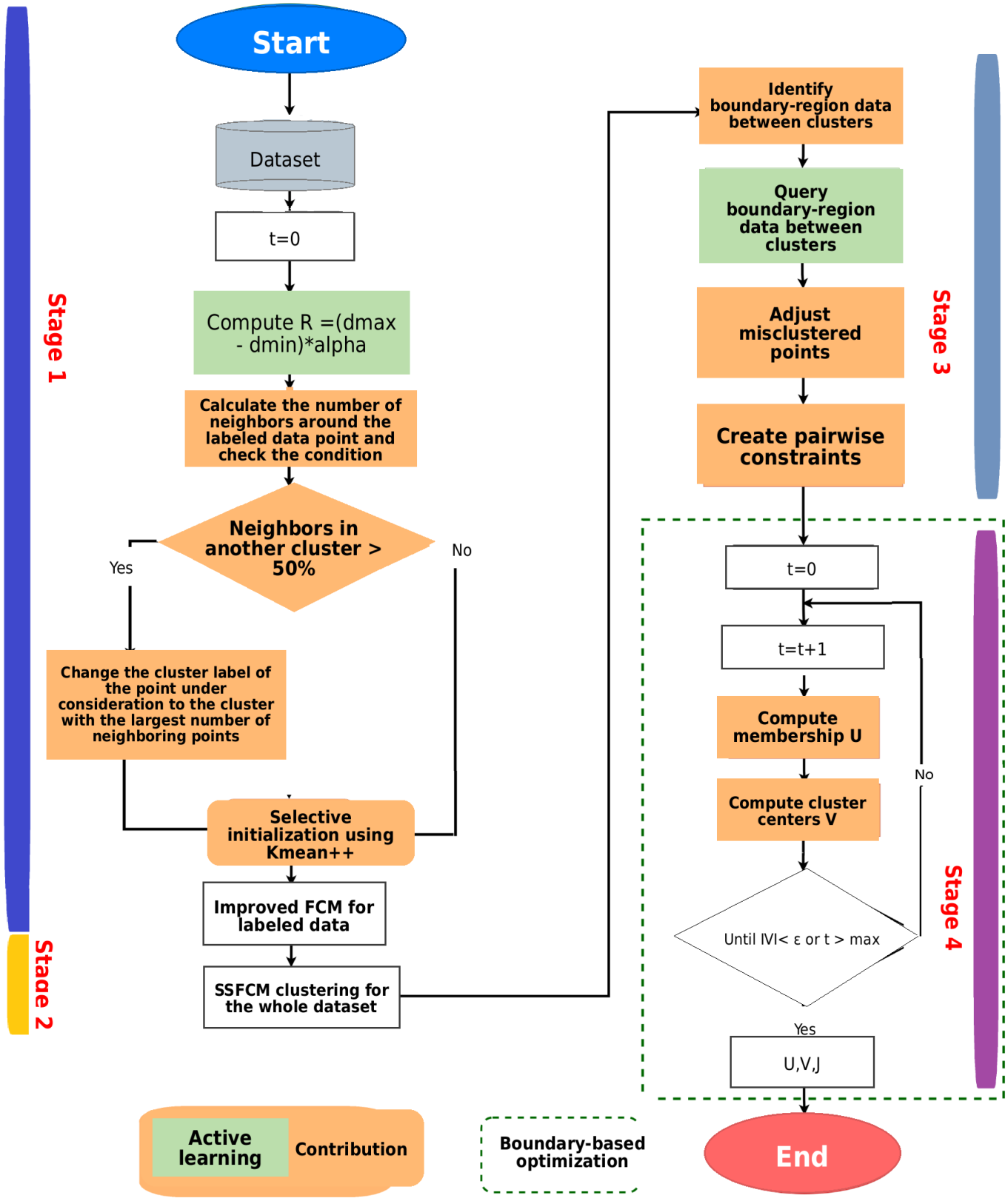

Figure 3.1: Active safe fuzzy semi-supervised clustering model with pairwise links based on cluster boundaries (AS3FCPC).

Next, the third stage focuses on refining the cluster boundaries. Boundary samples are detected based on a small membership difference between the two competing clusters; these samples are queried to the oracle to confirm or correct their labels,

from which must-link and cannot-link constraints are selectively generated, primarily in overlap regions where the model is most likely to make errors.

Finally, in the optimization and convergence stage, the algorithm minimizes an integrated objective function consisting of the fuzzy term, the semi-supervised term, and the pairwise constraint terms; the iterative process stops when the cluster-center deviation satisfies $\|V^{(t)} - V^{(t-1)}\| \leq \epsilon$ or when the maximum number of iterations is reached.

Objective Function

$$\begin{aligned}\min_{u,v}\ J(u,v) = &\sum_{i=1}^{N}\sum_{k=1}^{C} u_{ik}^2\, d_{ik}^2 + \alpha \sum_{i=1}^{L}\sum_{k=1}^{C}\big(u_{ik} - \bar{u}_{ik}\big)^2 \\ &+ \beta\left(\sum_{(x_i,x_j)\in\mathcal{M}}\sum_{k=1}^{C}\sum_{\substack{\ell=1\\ \ell\neq k}}^{C} u_{ik}\, u_{j\ell} + \sum_{(x_i,x_j)\in\mathcal{C}}\sum_{k=1}^{C} u_{ik}\, u_{jk}\right).\end{aligned} \tag{3.1}$$

subject to the constraints $\sum_{k=1}^{C} u_{ik} = 1,\ u_{ik}, \bar{u}_{ik} \in [0,1]$ and $d_{ik}^2 = \|x_i - v_k\|^2$.
Applying the Lagrange multiplier for each $k = 1, \ldots, N$, we obtain:

$$\mathcal{L}(U, V, \lambda) = J(U, V) + \sum_{i=1}^{N} \lambda_i \Big(\sum_{k=1}^{C} u_{ik} - 1 \Big) \tag{3.2}$$

The membership degree is then computed as follows:

$$u_{ik} = \frac{a_{ik} - \lambda_i}{2(d_{ik}^2 + \alpha)}, \qquad \lambda_i = \frac{\displaystyle\sum_{r=1}^{C} \frac{a_{ir}}{d_{ir}^2 + \alpha} - 2}{\displaystyle\sum_{r=1}^{C} \frac{1}{d_{ir}^2 + \alpha}}. \tag{3.3}$$

Cluster centers are updated by:

$$v_k = \frac{\displaystyle\sum_{i=1}^{N} u_{ik}^2 x_i}{\displaystyle\sum_{i=1}^{N} u_{ik}^2} \tag{3.4}$$

Active Safe Fuzzy Semi-Supervised Clustering with Pairwise Links Based on Cluster Boundaries - AS3FCPC

Algorithm 2 AS3FCPC

1: **Input:)** $X = \{x_i\}_{i=1}^{N}$, number of clusters C, number of labeled samples $L(< N)$, anchor matrix $\bar{U}$, convergence threshold ϵ, maximum number of iterations $maxStep$, constraint coefficient α.
2: **Stage 1 (Active learning + FCM on labeled data):** detect ambiguous samples using the neighborhood radius R, query/correct labels; run FCM on the labeled subset $\Rightarrow$ obtain initial U, V.
3: **Stage 2 (Iterative SSFCM):** update U, V over the whole dataset with the semi-supervised term $(u_{ik} - \bar{u}_{ik})^2 d_{ik}^2$; reinforce $\bar{U}$.
4: **Stage 3 (Boundary + pairwise constraints):** identify boundary samples via $|u_{i(k_1)} - u_{i(k_2)}| < \varepsilon$; query the oracle for boundary samples; selectively generate constraints $\mathcal{M}$ (must-link) and $\mathcal{C}$ (cannot-link).
5: **Stage 4 (Optimization until convergence):**
6: Set $t \leftarrow 0$.
7: **Repeat:**
8: $t \leftarrow t + 1$.
9: **Update membership matrix:** compute u_{ik} according to 3.3
10: **Update cluster centers:** compute v_k according to 3.4
11: **Until** $\|V^{(t)} - V^{(t-1)}\| \leq \epsilon$ **or** $t > maxStep$

The computational complexity of the **AS3FCPC** algorithm is:

$$\boxed{O\big(NC^2(I' + T) \; + \; N \log N\big)}$$

where I' denotes the number of TS3FCM iterations, T is the number of final refinement iterations, while FCM on the labeled subset with $L \ll N$ and the adjustment of $N_q \ll N$ samples are negligible ($O(LC^2I)$, $O(N_qC)$).
Since $T \ll I'$, the additional cost introduced by boundary refinement (active learning + pairwise constraints) is modest compared with the gain in accuracy and stability on overlapping data.
In summary, the computational cost remains of the same order as TS3FCM, but **AS3FCPC** achieves superior performance thanks to its directed boundary correction mechanism.

3.3 Experimental Results

Figure 3.2: Illustration of image segmentation for flooded-area detection

The dissertation evaluates AS3FCPC on three groups of datasets: (i) UCI datasets with diverse dimensionalities and class numbers; (ii) flood-region images with soft boundaries and high noise; and (iii) synthetic datasets specifically designed with strong overlap at cluster margins. Model performance is assessed using RI, F1, and NMI (the higher, the better), DB (the lower, the better), together with fuzzy measures PC, PE, and DB_fuzzy to reflect the sharpness of the membership matrix and cluster compactness in the fuzzy space.

Method	**IRIS**	**Wine**	**Australian**	**Spambase**	**Waveform**	**Breast**	**Haberman**	**Glass**	**Flood**
RI									
FCM	0.812	0.632	0.498	0.514	0.531	0.691	0.471	0.788	0.819
SSFCM	0.864	0.671	0.503	0.527	0.546	0.736	0.503	0.796	0.832
CS3FCM	0.887	0.693	0.513	0.562	0.512	0.748	0.518	0.804	0.873
TS3FCM	0.899	0.703	0.518	0.551	0.528	0.755	0.524	0.816	0.884
AFFC (2017)	0.928	0.715	0.563	0.582	0.566	0.761	0.536	0.822	0.869
AS3FCPC	**0.955**	**0.742**	**0.598**	**0.618**	**0.589**	**0.772**	**0.559**	**0.833**	**0.917**
F1-score									
FCM	0.793	0.513	0.497	0.502	0.529	0.632	0.512	0.525	0.788
SSFCM	0.800	0.573	0.504	0.521	0.516	0.675	0.539	0.537	0.812
CS3FCM	0.829	0.623	0.514	0.542	0.507	0.714	0.530	0.558	0.861
TS3FCM	0.839	0.633	0.508	0.532	0.505	0.726	0.552	0.584	0.867
AFFC (2017)	0.874	0.645	0.552	0.573	0.560	0.739	**0.566**	0.609	0.860
AS3FCPC	**0.911**	**0.659**	**0.576**	**0.597**	**0.574**	**0.761**	0.559	**0.618**	**0.894**
NMI									
FCM	0.658	0.362	0.442	0.457	0.486	0.413	0.453	0.603	0.732
SSFCM	0.702	0.398	0.466	0.492	0.502	0.452	0.468	0.614	0.746
CS3FCM	0.728	0.422	0.473	0.511	0.491	0.474	0.482	0.623	0.753
TS3FCM	0.740	0.434	0.478	0.522	0.499	0.485	0.493	0.647	0.766
AFFC (2017)	0.801	0.451	0.528	0.547	0.538	0.501	0.508	0.662	0.854
AS3FCPC	**0.833**	**0.472**	**0.563**	**0.594**	**0.563**	**0.527**	**0.517**	**0.684**	**0.883**
DB									
FCM	**0.902**	**0.826**	**1.143**	**1.609**	3.513	**0.731**	1.258	**0.629**	**1.252**
SSFCM	1.348	1.513	1.234	1.943	4.552	0.843	1.372	2.318	1.944
CS3FCM	2.785	3.028	4.313	4.034	5.684	1.932	**1.198**	1.347	3.524
TS3FCM	2.864	3.194	3.512	3.613	6.228	1.884	1.330	1.436	3.628
AFFC (2017)	1.349	0.983	2.816	2.098	2.432	0.942	1.417	2.334	2.528
AS3FCPC	1.222	0.937	2.275	1.843	**2.118**	0.963	1.528	2.485	1.327

Table 3.1: Overall performance comparison of the methods according to RI, F1, NMI, and DB

The experimental results show that AS3FCPC consistently achieves the best RI, F1, and NMI values on most datasets, outperforming FCM, SSFCM, CS3FCM, TS3FCM, and AFFC (2017). The improvements are particularly notable on IRIS, Waveform, Spambase, Glass (strong overlap), and the Flood image dataset (high noise). The fuzzy measures PC and PE also indicate that the memberships produced by AS3FCPC are sharper: PC is higher and PE is lower than those of most competing methods, confirming that the proposed model substantially reduces ambiguity at boundary points. Regarding the internal validity indices DB and DB_fuzzy, the method often achieves the best or highly competitive values among fuzzy semi-supervised clustering algorithms, although it is occasionally inferior to FCM.

Overall, the mechanism of active boundary correction combined with selective pairwise constraints reduces confusion between neighboring clusters, preserves the sep-

Method	IRIS	Wine	Australian	Spambase	Waveform	Breast	Haberman	Glass	Flood
PC									
FCM	0.780	0.720	0.611	0.605	0.626	0.741	0.591	0.691	0.705
SSFCM	0.811	0.736	0.629	0.618	0.640	0.765	0.615	0.706	0.728
CS3FCM	0.835	0.748	0.641	0.646	0.653	0.776	0.631	0.715	0.741
TS3FCM	0.851	0.761	0.655	0.659	0.670	0.782	0.646	0.725	0.755
AFFC (2017)	0.876	0.772	0.685	0.691	0.703	0.791	0.661	0.736	0.768
AS3FCPC	**0.910**	**0.795**	**0.706**	**0.721**	**0.726**	**0.805**	**0.686**	**0.761**	**0.811**
PE									
FCM	0.422	0.511	0.612	0.629	0.656	0.487	0.668	0.533	0.526
SSFCM	0.393	0.496	0.598	0.613	0.639	0.462	0.643	0.549	0.502
CS3FCM	0.371	0.487	0.583	0.598	0.617	0.454	0.628	0.523	0.476
TS3FCM	0.355	0.471	0.572	0.583	0.598	0.447	0.612	0.506	0.459
AFFC (2017)	0.329	0.460	0.556	0.566	0.584	0.440	0.589	0.492	0.433
AS3FCPC	**0.303**	**0.446**	**0.526**	**0.540**	**0.556**	**0.427**	**0.566**	**0.469**	**0.406**
DB_{fuzzy}									
FCM	**0.921**	**0.690**	**1.210**	**1.872**	2.413	**0.713**	1.246	**0.613**	**1.243**
SSFCM	1.430	1.513	1.384	2.105	4.563	1.028	1.383	2.318	1.973
CS3FCM	2.510	3.046	4.383	4.108	5.725	1.948	**1.213**	1.366	3.548
TS3FCM	2.620	3.216	3.546	3.658	6.266	1.918	1.343	1.453	3.673
AFFC (2017)	1.320	0.983	2.846	2.133	2.485	0.952	1.430	1.538	2.563
AS3FCPC	1.270	1.715	2.316	1.875	**2.158**	0.972	1.543	2.528	1.363

Table 3.2: Overall performance comparison of the methods according to PC, PE, and DB_{fuzzy}

aration structure, and improves external evaluation metrics. In scenarios requiring reliable label agreement under limited supervision and ambiguous boundaries (e.g., remote sensing image analysis), AS3FCPC demonstrates robust and effective performance.

3.4 CONCLUSION

In summary, AS3FCPC integrates fuzzy clustering, semi-supervised learning, and active learning to effectively handle problems with limited labels and unclear cluster boundaries. The algorithm is initialized in a stable manner through selective querying of uncertain data samples and the propagation of pairwise constraints, thereby shaping the cluster structure from the early stage. In subsequent optimization steps, the model continues to adjust memberships and update cluster centers under must-link/cannot-link constraints, helping stabilize boundaries, reduce confusion between neighboring clusters, and increase cluster separability, especially in highly overlapping regions.

Experimental results on multiple datasets show that the external indices RI, F1, and NMI of AS3FCPC consistently attain higher values than those of comparison methods, while the fuzzy measures PC and PE reach favorable levels, reflecting a sharper fuzzy partition and well-controlled uncertainty. At the same time, the internal indices DB and DB_fuzzy are generally improved or maintained at competitive levels compared with other fuzzy semi-supervised clustering algorithms, indicating that the obtained cluster structure is not only consistent with true labels but also compact and stable. Owing to these characteristics, AS3FCPC is particularly suitable for practical applications such as flood-area delineation in satellite imagery and has been presented in work CT3.

CONCLUSION

A. Main results of the dissertation

The dissertation investigates fuzzy semi-supervised clustering, synthesizing the theoretical foundations of fuzzy sets, FCM, semi-supervised learning, active learning, and safe learning, and accordingly proposes two principal methods:

1. **ASSFBC**: refines cluster boundary regions through an active querying mechanism for informative samples, thereby improving membership assignment accuracy and clustering reliability.
2. **AS3FCPC**: combines fuzzy clustering, safe semi-supervised learning, and active learning; simultaneously exploits labeled and unlabeled data from the initialization stage, and uses pairwise constraints to stabilize cluster centers and cluster boundaries.

B. Novel contributions of the dissertation

- Integrating active learning into fuzzy semi-supervised clustering with a *boundary-focused* strategy, thereby improving clustering quality under a limited labeling budget.
- Proposing the **AS3FCPC** model, which harmonizes fuzzy clustering, safe semi-supervised learning, and active learning, while effectively exploiting pairwise constraints to enhance both accuracy and stability.
- Conducting comprehensive experimental validation, with superior results in terms of accuracy, reliability, and overall model effectiveness.

C. Future research directions

Future research will focus on optimizing ASSFBC and AS3FCPC for large-scale data, while further exploiting domain-specific context (such as healthcare, remote sensing, and social networks) to improve the adaptability of the proposed models. In addition, it is necessary to develop dynamic active learning strategies suitable for streaming data and noisy environments, as well as to integrate deep learning techniques in order to build more powerful active fuzzy semi-supervised clustering models.

LIST OF THE PUBLICATIONS RELATED TO THE DISSERTATION

1. Duong Tien Dung, Nguyen Long Giang, Hoang Viet Long, Tran Manh Tuan, Luong Thi Hong Lan, Dinh Thu Khanh (2021), "An advancement in active fuzzy semi-supervised clustering", Proceedings of the 24th National Conference on Information and Communication Technology (VNICT 2021).
2. Duong Tien Dung, Ha Hai Nam, Nguyen Long Giang, Luong Thi Hong Lan (2025), "An Innovative Semi-Supervised Fuzzy Clustering Technique Using Cluster Boundaries", Computers, Materials & Continua 2025, 85(3), 5341-5357 (SCIE Q3: IF 1.7; Scopus CiteScore: 6.1). https://doi.org/10.32604/cmc.2025.068299.
3. Duong Tien Dung, Ha Hai Nam, Nguyen Long Giang, Luong Thi Hong Lan (2025), "An Active Safe Semi-Supervised Fuzzy Clustering with Pairwise Constraints Based on Cluster Boundary", Computers, Materials & Continua 2025, 85(3), 5625-5642 (SCIE Q3: IF 1.7; Scopus CiteScore: 6.1). https://doi.org/10.32604/cmc.2025.069636