MINISTRY OF EDUCATION AND TRAINING

VIETNAM ACADEMY OF SIENCE AND TECHNOLOGY

GRADUATE UNIVERSITY OF SIENCE AND TECHNOLOGY

Pham Minh Ngoc Ha

SOME NOVEL METHODS TO IMPROVE THE EFFICIENCY OF CALCULATING THE REDUCT ON FUZZY APPROXIMATION SPACES

SUMMARY OF DISSERTATION ON: INFORMATION SYSTEM
Code: 9 48 01 04

Ha Noi – 2023

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

Supervisors:

1. Supervisor 1: Associate Professor . Dr. Nguyen Long Giang, Information of Technology Institute, Vietnam Academy of Science and Technology, Ha Noi, Viet Nam

2. Supervisor 2: Dr. Nguyen Manh Hung, Military Technical Academy, Hanoi, Vietnam

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 .....

The dissertation can be found at:

.....1. Graduate University of Science and Technology Library

2. National Library of Vietnam

INTRODUCTION

Urgency of the Dissertation Topic

Feature selection, also known as attribute selection, is an important step in data analysis and statistical machine learning. This process involves selecting a subset of relevant attributes from the original attribute set in such a way that important information is maximally preserved. Feature selection offers many significant benefits: 1) reducing computational complexity, 2) improving model interpretability, and 3) enhancing predictive performance. The main goal is to find a subset of features, from the original set, that still ensures the retention of information or accurate decision-making capability. Important applications of feature selection are widespread in areas such as pattern recognition and data mining, including text classification [1], [2], image processing [3]–[5], and speech processing [6]–[9].

In 1982, Pawlak introduced the Rough Set (RS) theory [10], which has been highly regarded by the scientific community for its ability to analyze data under incomplete and inconsistent conditions. Thanks to this ability, RS-based attribute selection has attracted significant interest among researchers in the field of rough set theory for many years [4], [11]–[16]. Based on the concept of Approximation Space (AP) in RS, many measures have been proposed to define reducts and support attribute selection. Recent studies show that traditional RS-based attribute reduction methods yield remarkable results in reducing the number of attributes while still preserving the classification ability of decision tables [1], [8], [14]–[17].

However, traditional rough sets are mainly suitable for decision tables with discrete value domains [18]. Therefore, it is necessary to discretize the data in numerical decision tables (with continuous value domains) before performing attribute selection. This process incurs additional computational cost, may compromise the natural form of the data, and poses a risk of losing important information. To address these limitations, researchers have proposed extending RS to fuzzy approximation spaces, leading to the Fuzzy Rough Set (FRS) model [19]–[21], and further to intuitionistic fuzzy approximation spaces, resulting in the Intuitionistic Fuzzy Rough Set (IFRS) model [22]. These models allow the development of attribute selection methods directly on original decision tables without the need for discretization.

The fuzzy approximation space of FRS uses the concept of similarity relation instead of equivalence relation to construct relational spaces among objects. As a result, relationships in the fuzzy approximation space become more flexible than traditional equivalence-based relationships. The degree of relationship among objects is represented by values in the range $[0, 1]$ instead of just 0 or 1 as in traditional rough sets. Currently, the development of attribute selection methods based on measures built in fuzzy approximation spaces is very active, with several notable measures proposed, including FPOS measures [20], [21], [23]–[28], FIE measures [4], [29]–[32], and FD measures [6], [11], [33]–[35]. In Vietnam, the dissertation by Dr. Nguyễn Văn Thiện extended several measures in fuzzy approximation space for attribute reduction in numerical decision tables.

Although studies have shown that attribute selection methods based on FRS measures work effectively with numerical datasets, their performance may degrade when applied to highly sensitive datasets with high misclassification rates (i.e., noisy data). Therefore, the Variable Precision Fuzzy Rough Sets (VPFRS) model [19], [31], [36]–[50] and measures developed in the intuitionistic fuzzy approximation space of the IFRS model [13], [20], [22], [41], [46], [51]–[54] have been proposed to address this issue.

In the VPFRS approach, adjusting the components in the lower approximation set affects

the positive region of the attribute, leading to changes in the attribute's dependency degree. In contrast, IFRS-based measures entirely depend on the intuitionistic fuzzy approximation space. In this space, each element represents degrees of similarity and dissimilarity between two objects. Thus, the intuitionistic fuzzy approximation space describes multi-dimensional relationships among objects more comprehensively than FRS [52]. Research [51] has shown that the IFRS approach can improve reduct quality in noisy datasets, though it incurs higher computational costs (twice that of FRS). In Vietnam, Dr. Trần Thanh Đại's dissertation proposed the Intuitionistic Fuzzy Distance (IFD) measure for attribute selection in numerical decision tables containing noise. However, its computational cost remains high due to the complexity of defining membership and non-membership functions for the AP. Hence, the *first research objective* of this dissertation is to study and extend the VPFRS model to improve the efficiency of approximation set computations. This objective belongs to the *group of attribute selection methods for static decision tables*, meaning decision tables whose content does not change over time.

In practice, machine learning applications often require model updates to adapt to changes in data over time. Therefore, developing efficient attribute selection methods for updated data is an urgent need [55]. So far, many incremental computation methods have been proposed to efficiently update reducts [3], [6], [55]–[92]. These incremental techniques only evaluate new information and combine it with previous results to update reducts.

There are three main scenarios of data change: changes in attribute sets [55], [56], [58], [60], [61], changes in object sets [3], [6], [64], [69], [70], [74], [78]–[80], and changes in the content of objects [3].

In Vietnam, the doctoral dissertation by Nguyễn Bá Quảng proposed an incremental computation method based on distance measures, built on the monotonicity of union and intersection operations between two sets. Recently, Yang and colleagues proposed an incremental method using an information-theoretic granular measure approach [70]. Unlike distance-based measures, information-theoretic granular measures rely on the coarseness and fineness of partitions, leading to simpler formula construction and faster computation. However, Zhang et al.'s work has only developed this measure in the crisp approximation space, not in the fuzzy approximation space.

Therefore, the *second research objective* of this dissertation is to study and extend the information-theoretic granular measure in fuzzy approximation space and apply it to develop an attribute update method for numerical decision tables with changing object sets. This second objective belongs to the *group of attribute selection methods for decision tables with changing objects*.

Dưới đây là phần tiếp theo của bản dịch tiếng Anh của chương "Giới thiệu" trong luận văn, tiếp nối sau phần "Urgency of the Dissertation Topic":

"'latex **Objectives of the Dissertation**

The dissertation focuses on proposing new attribute selection methods in fuzzy approximation spaces for numerical decision tables in two groups of problems: (1) static decision tables, and (2) decision tables with changing object sets. These methods are based on extending the VPFRS model and the fuzzy information-theoretic granular measure. Specifically, the dissertation aims to address the following main objectives:

- **Objective 1**: Research and extend the VPFRS model to improve the performance of approximation set computations in numerical decision tables.
- **Objective 2**: Research and extend the fuzzy information-theoretic granular measure to propose an attribute update method for numerical decision tables with changing object sets.

Scope of the Dissertation

The dissertation focuses on studying and proposing attribute selection methods for numerical decision tables (tables with continuous attribute values) based on fuzzy approximation spaces. The study emphasizes two main directions:

- Attribute selection for static decision tables, where the content of the table remains unchanged.
- Attribute update for decision tables with changing object sets, using incremental computation.

The scope of the dissertation does not cover attribute selection methods for tables with changing attribute sets or changing object values.

Research Methods

The research methodology of the dissertation includes:

- Analyzing and extending existing theoretical models (VPFRS, fuzzy granular measures) to enhance attribute selection capabilities.
- Constructing new evaluation measures and attribute selection algorithms in fuzzy approximation spaces.
- Designing and implementing algorithms for static and dynamic datasets.
- Conducting experiments on standard datasets from machine learning repositories to evaluate the performance of the proposed methods.

Scientific Contributions of the Dissertation

The dissertation makes the following main contributions:

- Proposing an improved fuzzy rough set model based on the variable precision concept (VPFRS) to enhance attribute selection effectiveness in numerical decision tables.
- Extending the fuzzy information-theoretic granular measure for attribute selection and proposing an attribute update method for decision tables with changing object sets.
- Developing new algorithms that achieve competitive results in terms of accuracy and efficiency when compared to existing methods on benchmark datasets.

Structure of the Dissertation

The dissertation is structured as follows:

- **Chapter 1**: Presents an overview of rough sets, fuzzy rough sets, and related theories in attribute selection.
- **Chapter 2**: Analyzes and proposes a VPFRS-based model and related measures for static decision tables.
- **Chapter 3**: Studies and proposes a fuzzy information-theoretic granular measure and incremental attribute update method for changing object sets.
- **Conclusion**: Summarizes the main contributions and outlines future research directions.

CHAPTER 1. Overview of Attribute Reduction Methods in Fuzzy Approximation Spaces

Chapter 1 of the dissertation focuses on providing an overview of the attribute selection problem (also known as attribute reduction or feature selection) and the foundational knowledge necessary for subsequent chapters [1]. This chapter presents an overview of attribute selection methods in fuzzy approximation spaces, focusing on two main approaches: improving classification accuracy on noisy numerical decision tables and incremental computation on numerical decision tables updated with new objects [1, 2].

1.1. Introduction

The introduction highlights the urgency of the dissertation topic, emphasizing that attribute selection is a critical step in data preprocessing for data analysis and machine learning tasks, particularly in the context of increasingly large and complex datasets [3, 4]. This process involves selecting a subset of relevant attributes from the original attribute set, preserving the information and classification capability of the original data while reducing computational complexity, improving model interpretability, and enhancing predictive performance [4, 5].

1.1.1. Overview of the Attribute Selection Problem Using the Rough Set Approach

The dissertation revisits Pawlak's introduction of the Rough Set (RS) theory in 1982, a powerful tool for analyzing incomplete and inconsistent data [6, 7]. Leveraging this capability, attribute selection using the RS approach has garnered widespread attention [5, 6]. Unlike dimensionality reduction methods such as PCA or SVD, this approach directly selects attributes without altering the natural characteristics of the reduced data [5].

The attribute selection problem using the rough set approach involves the following key steps [8]:

- Defining the search space: The set of all possible subsets of the original attribute set.
- Constructing an attribute evaluation function: Using measures based on the concept of approximation spaces in RS to assess the importance of individual attributes or attribute subsets. Common measures include positive region measure, dependency measure, and discrimination measure.
- Applying search strategies: Employing various search algorithms (e.g., exhaustive search, greedy search, genetic algorithms) to identify the optimal attribute subset based on the constructed evaluation function.
- Validating the correctness of the subset: Ensuring that the selected attribute subset preserves the information and classification capability of the original data.

The sht dissertation also discusses three main approaches to attribute selection [8]:

- Filter approach: Relies on the inherent properties of the data (e.g., information, correlation) to evaluate and select attributes independently of any specific machine learning algorithm [8, 9]. Common strategies include adding important attributes (Figure 1.1 [10]) or removing redundant attributes (Figure 1.2 [11]).

- Wrapper approach: Uses a specific machine learning algorithm as the evaluation function. The search for the optimal attribute subset is performed by training and evaluating the performance of the machine learning algorithm on different attribute subsets (Figure 1.3 [12]). This approach typically offers higher accuracy but incurs greater computational complexity, especially with exhaustive search strategies [9].
- Hybrid filter-wrapper approach: Combines the advantages of both filter (fast computation) and wrapper (high accuracy) methods [13].

1.1.2. Some Applications of the Attribute Selection Problem

This section discusses the importance and wide applicability of attribute selection across various domains [13]:

- Data classification: Improves the performance of classification models by selecting the most discriminative attributes and reducing the risk of overfitting [14].
- Data regression: Identifies the most influential predictive factors and simplifies the model while maintaining predictive capability [14].
- Data clustering: Enhances the efficiency of clustering algorithms by eliminating irrelevant or redundant attributes, leading to more meaningful clusters [14].
- Information security: Filters out harmful or risky features, such as in web spam detection (reducing computation time and cost [15, 16]) and malware detection (selecting critical API functions [16]).

1.2. Dependency Measures of Attributes in Fuzzy Approximation Spaces

To address the limitations of traditional rough sets with continuous-valued data, researchers have developed the Fuzzy Rough Set (FRS) model in fuzzy approximation spaces [17, 18]. In FRS, the degree of relationship between objects is represented by values in the interval [19], enabling better handling of ambiguity and uncertainty in numerical data [20]. Several measures have been proposed to define reducts and support attribute selection in fuzzy approximation spaces [20, 21]. These measures are typically built upon extensions of traditional rough set models [21].

1.2.1. Fuzzy Positive Region Measure

The positive region measure is a statistical measure based on the values of lower approximation sets in rough set theory [21]. The fuzzy positive region measure extends this concept to fuzzy approximation spaces, representing the dependency of decision attributes on conditional attributes through the statistical positive region values [21, 22]. This measure enables the evaluation of the consistency of a decision table (a table is considered consistent if the dependency of the decision attribute is maximized, i.e., equal to 1) [22]. Table 1.1 [23] summarizes studies related to the development of fuzzy positive region measures for different data types and approximation space models [22, 24].

1.2.2. Fuzzy Entropy Measure

Shannon's entropy measure is widely used in machine learning, particularly in decision trees, to determine the amount of information required to classify each element [24]. In fuzzy approximation spaces, the entropy measure is extended to evaluate the information gain of each attribute, serving as a criterion for ranking their importance in decision tables [25]. Table 1.2 [26] lists studies on fuzzy entropy measures across different data types and approximation spaces [25, 27].

1.2.3. Information Granularity Measure

The information granularity measure is a key concept in granular computing [27]. In traditional rough sets, the granularity measure is based on statistical values of a partition or cover, reflecting the coarseness or fineness of the partition [27]. The dependency measure based on granular computing is used to evaluate attribute dependency in decision tables through the coarseness or fineness of dependent information granules [27, 28]. This measure has been extended to fuzzy approximation spaces to support attribute selection on numerical data [28]. Table 1.3 [26] summarizes studies related to fuzzy information granularity measures [28, 29].

1.3. Attribute Selection Methods for Improving Accuracy in Fuzzy Approximation Spaces

The growth of big data has increased instances of missing, incomplete, and inconsistent data, often containing noisy attributes [29]. Identifying and eliminating these attributes is crucial for reducing training time and improving model accuracy [29, 30].

1.3.1. Related Studies

For decision tables with discrete and incomplete values, the Variable Precision Rough Set (VPRS) model has been proposed to adjust the components of approximation sets based on membership degrees, enabling better noise handling [30, 31].

For numerical decision tables, the VPRS model has been extended to the Variable Precision Fuzzy Rough Set (VPFRS) in fuzzy approximation spaces [31]. In VPFRS, lower and upper approximation operations can be adjusted to vary the precision of approximation sets, affecting attribute dependency [31, 32]. However, this method heavily relies on the selection of the adjustment threshold (β), which varies across datasets and requires significant expertise [32].

Another approach involves using Intuitionistic Fuzzy Rough Sets (IFRS) in fuzzy approximation spaces [32, 33]. IFRS imposes stricter constraints based on membership and non-membership functions, describing object relationships in a more multidimensional manner than FRS [33-35]. IFRS has been shown to improve reduct quality on noisy data but incurs higher computational costs [35]. Constructing independent membership and non-membership functions also poses a challenge [33].

Table 1.4 [17] summarizes studies related to attribute selection methods for improving classification accuracy, including VPRS, VPFRS, and IFRS approaches [36, 37].

1.3.2. Existing Issues

Recent studies on attribute selection methods for improving classification accuracy, particularly those based on VPFRS, are limited in terms of computational efficiency [38, 39]. Evaluating all objects in a decision table during precision adjustment leads to high computational costs

[39]. The dissertation aims to improve computational efficiency by optimizing approximation operations based on the fundamental properties of variable precision fuzzy rough sets [36, 40].

1.4. Incremental Attribute Selection Methods in Fuzzy Approximation Spaces

In many real-world applications, data is continuously updated and changed over time, requiring classification models to be adjusted accordingly [40]. Incremental attribute selection plays a critical role in reducing data preprocessing time during changes, thereby improving the efficiency of model updates [40, 41].

1.4.1. Related Studies

Attribute selection methods based on measures derived from lower approximation sets are typically suitable for static decision tables [41]. However, in practice, decision tables are frequently updated, requiring efficient computation based solely on new information [41, 42].

Incremental computation methods have been proposed to efficiently update reducts when data changes [43]. There are three main data change scenarios [42, 43]:

- Object set changes: Adding or removing data samples [42, 44]. Numerous studies have focused on developing incremental formulas and algorithms for updating reducts in this scenario, using measures such as information granularity [44], information entropy [44], and functional dependency [44]. Table 1.5 [20] summarizes incremental methods for object set changes [42, 44].
- Attribute set changes: Adding or removing attributes [42, 45]. Studies in this area also focus on developing efficient methods for updating reducts when the attribute set changes, often using the information granularity measure [45]. Table 1.6 [46] lists incremental methods for attribute set changes [45].
- Attribute value changes.
- Mixed changes: Combinations of multiple change types [42, 45].

1.4.2. Existing Issues

Incremental computation methods based on knowledge distance measures often have complex formulas, making them difficult to prove and verify [45, 47]. In contrast, incremental computation formulas based on knowledge information granularity measures have simpler structures and are easier to prove in cases of object addition or removal [47]. However, current knowledge information granularity measures are primarily developed for crisp approximation spaces and have not been extended to fuzzy approximation spaces [48]. The dissertation aims to study and extend the knowledge information granularity measure to fuzzy approximation spaces, applying it to develop attribute update methods for numerical decision tables with object changes [48, 49].

1.5. Foundational Knowledge

This section provides basic definitions and concepts that serve as the foundation for subsequent chapters [7].

1.5.1. Traditional Rough Sets

Pawlak's rough set model is a mathematical approach for handling incomplete, inaccurate, and ambiguous data, based on the concept of set approximation [47].

- Information System (IS): Defined as a pair $S = (U, A)$, where U is a finite set of objects and A is a finite set of attributes [50]. Each attribute $a \in A$ is a function $a : U \to V_a$, assigning a value from the value set V_a to each object.
- Decision Table (DT): An information system $NDT = (U, C \cup D)$, where C is the set of conditional attributes and D is the set of decision attributes [51].
- Indiscernibility Relation: For a subset of attributes $B \subseteq A$, the indiscernibility relation $IND(B)$ is an equivalence relation on U defined as follows: $(x, y) \in IND(B)$ if and only if $a(x) = a(y)$ for all $a \in B$. This relation partitions U into equivalence classes $[x]_B = \{y \in U | (x, y) \in IND(B)\}$ [50].
- Lower Approximation: For $X \subseteq U$, the lower approximation of X with respect to B, denoted $\underline{B}X$, is the set of all objects that certainly belong to X based on the information in B: $\underline{B}X = \{x \in U | [x]_B \subseteq X\}$ [50].
- Upper Approximation: For $X \subseteq U$, the upper approximation of X with respect to B, denoted $\overline{B}X$, is the set of all objects that possibly belong to X based on the information in B: $\overline{B}MulX = \{x \in U | [x]_B \cap X \neq \emptyset\}$ [50, 52].
- Boundary Region: The boundary region of X with respect to B, denoted $BN_B(X)$, contains objects that cannot be definitively classified as belonging or not belonging to X: $BN_B(X) = \overline{B}X - \underline{B}X$ [52].
- Rough Set: A set X is called a rough set if $\underline{B}X \neq \overline{B}X$, i.e., the boundary region is non-empty [52]. The rough set theory model is illustrated in Figure 1.4 [11, 52].

1.5.2. Traditional Fuzzy Sets

Fuzzy set theory, introduced by Zadeh, addresses ambiguity in data [53].

- Fuzzy Set: A fuzzy set A in a universe U is defined by a membership function $\mu_A : U \to [0, 1]$, where $\mu_A(x)$ represents the degree of membership of element x in the fuzzy set A [43, 53].
- Basic Fuzzy Set Operators: Table 1.8 [3] describes basic operators such as complement, union, and intersection of fuzzy sets, with different T-norm (e.g., min, product) and T-conorm (e.g., max, algebraic sum) operations [43].

1.5.3. Fuzzy Approximation Spaces

A fuzzy approximation space combines an object space with a fuzzy equivalence relation [43].

- Fuzzy Equivalence Relation: A fuzzy relation $R : U \times U \to [0, 1]$ is called a fuzzy equivalence relation if it satisfies three properties [49, 54, 55]:

1. Reflexivity: $R(x,x) = 1$ for all $x \in U$.
2. Symmetry: $R(x,y) = R(y,x)$ for all $x,y \in U$.
3. Min-Transitivity: $R(x,z) \geq \min\{R(x,y), R(y,z)\}$ for all $x,y,z \in U$.

- Fuzzy Equivalence Relation Matrix of an Attribute: For a decision table $NDT = (U,C,D)$, the relation matrix $S(a,U)$ of an attribute $a \in C$ represents the degree of similarity between objects based on attribute a. An example of constructing a relation matrix from numerical data is presented in [55].
- Fuzzy Equivalence Relation Matrix of an Attribute Set: For $P \subseteq C$, the relation matrix $S(P,U)$ is the intersection of the relation matrices of each attribute in P, typically using the min operation: $s_{P_{ij}} = \min\{s_{a_{ij}} : a \in P\}$ [55].
- Fuzzy Partition: For a fuzzy equivalence relation R on U, the fuzzy partition of R on U is $U/R = \{[x]_R : x \in U\}$, where $[x]_R$ is the fuzzy equivalence class of x [56]. Similarly, the fuzzy partition of U with respect to an attribute $a \in C$ is U/R_a [56], and with respect to an attribute set $P \subseteq C$ is U/R_P [56, 57].

1.5.4. Fuzzy Rough Sets

Fuzzy rough sets are an extension of traditional rough sets in fuzzy approximation spaces [18].

- For a decision table $NDT = (U,C,D,f)$ and a fuzzy equivalence relation R on U, for any fuzzy set $A \in F(U)$, the membership degree of an object $x \in U$ to the fuzzy lower approximation $\underline{R}A(x)$ and fuzzy upper approximation $\overline{R}A(x)$ of A with respect to R is determined as follows [58]:
  - $\underline{R}A(x) = \inf_{y \in U} I(R(x,y), A(y))$
  - $\overline{R}A(x) = \sup_{y \in U} \min(R(x,y), A(y))$

  where I is a fuzzy implication function. An example of calculating fuzzy approximations is illustrated in [58].

1.6. Process of Developing and Evaluating Feature Selection Algorithms

This section describes the general process for developing and evaluating feature selection algorithms, applicable to both traditional and extended rough sets [59].

1.6.1. Steps for Developing Feature Selection Methods

The basic steps include [59]:

- Developing methods for evaluating, classifying, and selecting attributes: Using approaches such as discrimination matrices (classifying attributes based on discrimination levels) or measure-based approaches (evaluating dependency, information content, or information granule size) [59].
- Constructing reducts: Based on necessary conditions (preserving classification capability) and sufficient conditions (minimizing the number of attributes) defined [59].

- Designing algorithms: Using filter methods (addition or removal strategies), wrapper methods (combined with evolutionary algorithms), or hybrid methods to search for reducts [60]. Algorithm 1.1 [60] describes a general model for most reduction algorithms.

1.6.2. Data and Data Standardization

When building machine learning models, data standardization is a critical step to ensure model performance and accuracy [61, 62]. Basic standardization tasks include [61, 62]:

- Unit standardization: Converting units to the same value domain or measurement unit [61].
- Handling missing values: Imputing or removing missing values [61].
- Time standardization: Synchronizing temporal information [62].
- Scale standardization: Ensuring variables have the same scale to avoid disproportionate impacts on the model [62].

1.6.3. Classification Models and Evaluation Methods

The dissertation discusses the use of standard data classification models for numerical data to evaluate the classification capability of obtained reducts [63]. Some common models include [64]:

- K-Nearest Neighbors (k-NN)
- Support Vector Machines (SVM)

To evaluate the performance of classification models, commonly used metrics include [63, 64]:

- Accuracy: The proportion of correctly classified instances.
- Precision: The proportion of instances predicted as positive that are actually positive.
- Recall: The proportion of actual positive instances correctly predicted as positive.
- F1-score: The harmonic mean of precision and recall.

Table 1.9 [3] illustrates the binary confusion matrix, the basis for calculating these metrics [63].

The k-fold cross-validation method is a key technique for objectively and accurately evaluating models [63, 65, 66]. This process involves dividing the data into k parts, repeating the training and evaluation process k times, each time using one part as the test set and the remaining k-1 parts as the training set [66, 67]. The average accuracy is calculated after k iterations [67]. This method helps mitigate overfitting and underfitting [68]. The dissertation specifically mentions the 10-fold cross-validation method [68].

1.7. Conclusion of Chapter 1

Chapter 1 has provided an overview of the feature selection problem, its significance, and its role in machine learning and data mining [69]. It reviews related studies on two main groups of methods: improving classification accuracy and incremental feature selection [69].

Analysis of recent studies indicates that methods for improving classification accuracy based on VPFRS are limited in computational efficiency due to the need to evaluate all objects [69, 70]. The IFRS approach has higher complexity, and constructing suitable membership/non-membership functions is challenging [70, 71]. Therefore, the dissertation plans for Chapter 2 to study the extension of the VPFRS model to improve classification accuracy and computational efficiency [71].

For incremental feature selection methods when decision tables change in terms of objects, the knowledge information granularity measure has simpler formulas compared to distance-based measures [72]. However, this measure has not been extended to fuzzy approximation spaces [72, 73]. Thus, the dissertation plans for Chapter 3 to study and extend the knowledge information granularity measure in fuzzy approximation spaces to develop incremental computation formulas [73].

CHAPTER 2. Attribute Reduction Using an Extended Variable Precision Fuzzy Rough Set Approach in Fuzzy Approximation Spaces

Chapter 2 of the dissertation focuses on the problem of **attribute selection to improve classification accuracy for noisy numerical decision tables** [1]. Stemming from the **computational time limitations of methods based on the Variable Precision Fuzzy Rough Set (VPFRS) model** and the complexity of Intuitionistic Fuzzy Rough Set (IFRS) approximation spaces, this chapter proposes a **new approach by extending the VPFRS model in fuzzy approximation spaces** [1, 2]. The primary objective is to **improve computational efficiency during attribute reduction** while **enhancing classification accuracy on noisy datasets** [2]. The proposed method, named VPOFRS (Variable Precision Optimized Fuzzy Rough Set), promises to significantly reduce computational time compared to previous methods [2]. The research results have been published in works [CT2, CT3] [2].

2.1. Preliminary Knowledge

This section reviews key concepts that form the foundation for the proposals in this chapter. First is the **Variable Precision Rough Set (VPRS)** model, proposed by Ziarko in 1993, applied to attribute reduction to improve classification accuracy for incomplete classification decision tables [3]. Next is the **Variable Precision Fuzzy Rough Set (VPFRS)** model, introduced by Zhao in 2009 [4], an extension of VPRS in fuzzy approximation spaces, proven effective for attribute reduction on numerical datasets [3]. However, the approximation set computation formulas in VPFRS (formulas 2.8 and 2.9 in the dissertation) are not yet optimized, as they require evaluating all objects [5].

2.2. Proposed Improvement to the Variable Precision Fuzzy Rough Set Model

To address the limitations of the existing approximation set computation formulas in VPFRS, this section presents an improved method for operations related to object membership degrees and determining object boundaries [6].

2.2.1. Proposed Fuzzy Approximation Space

Definition 2.7 (Fuzzy Equivalence Relation): Let U be a non-empty set of objects. A fuzzy relation $R : U \times U \to [0,1]$ is called a fuzzy equivalence relation if it satisfies three properties:

- **Reflexivity**: $R(x,x) = 1, \forall x \in U$.
- **Symmetry**: $R(x,y) = R(y,x), \forall x,y \in U$.
- **T-norm Transitivity**: $T(R(x,y),R(y,z)) \leq R(x,z), \forall x,y,z \in U$, where T is a T-norm. Typically, $T(a,b) = \min(a,b)$ is used [5].

Definition 2.8 (Fuzzy Approximation Space): A pair (U,R) is called a fuzzy approximation space, where U is a non-empty set of objects and R is a fuzzy equivalence relation on U [5, 8].

Definition 2.9 (Fuzzy Partition): For a decision table $NDT = (U,C,D)$, and R a fuzzy equivalence relation on U, the fuzzy partition of R on U is defined as follows:

$$U/R = \{[x]_R : x \in U\}$$

where $[x]_R$ is the fuzzy equivalence class containing x, defined by $[x]_R(y) = R(x,y), \forall y \in U$ [8].

Definition 2.10 (Fuzzy Partition of an Attribute): For a decision table $NDT = (U,C,D)$, and R a fuzzy equivalence relation on U, the fuzzy partition of U with respect to an attribute $a \in C$ on R is denoted by U/R_a and is determined by formula (2.12) [8].

Definition 2.11 (Fuzzy Partition of an Attribute Set): For a decision table $NDT = (U,C,D)$ and $P \subseteq C$ a subset of conditional attributes, the fuzzy equivalence relation R_P is defined as the intersection of the fuzzy equivalence relations corresponding to each attribute in P:

$$R_P(x,y) = \min_{a \in P}\{R_a(x,y)\}, \forall x,y \in U$$

The fuzzy partition of U with respect to the attribute set P on R is denoted by U/R_P and is defined by $U/R_P = \{[x]_{R_P} : x \in U\}$ [8]. Propositions 2.1 and 2.2 in the dissertation prove properties related to the relationship between fuzzy partitions under attribute set inclusion [9].

Definition 2.12 (Finest Fuzzy Partition): For a decision table $NDT = (U,C,D)$ and P a fuzzy partition of U with $P \subseteq C$, P is called the finest partition if for every $x \in U, [x]_P = \{x\}$ [9].

Definition 2.13 (Coarsest Fuzzy Partition): For a decision table $NDT = (U,C,D)$ and Q a fuzzy partition of U with $Q \subseteq C$, Q is called the coarsest partition if for every $x \in U, [x]_Q = U$ [9].

2.2.2. Proposed Variable Precision Fuzzy Rough Set Model

Based on the limitations of the existing approximation set computation formulas (2.8 and 2.9), this section presents an improved method for calculating the membership degree of $x \in U$ and determining the boundaries of $y \in U$ [6].

Definition 2.14 (Improved Membership Degree of $x \in U$): Let U be a non-empty set of objects, for all $x,y \in U, A \subseteq U$.

- The β-membership of x to the lower approximation of A with respect to relation R on U is defined as follows:

$$R_\beta A(x) = \min\{\inf_{y \in U} I(R(x,y),A(y)), \beta\}$$

where $I(a,b) = 1 - a + a \cdot b$ is the Lukasiewicz implication function.

- The β-membership of x to the upper approximation of A with respect to relation R on U is defined as follows:

$$\overline{R}_\beta A(x) = \max\{\sup_{y \in U} \min(R(x,y), A(y)), 1-\beta\}$$

The thresholds $\beta \in [0,1]$ allow adjustment of the pessimistic/optimistic degree in determining approximation sets [6].

Proposition 2.3 (Improved Boundary of $y \in U$): For a decision table $NDT = (U,C,D)$, with an attribute set $B \subseteq C$ and a decision class D_k, the β-precision fuzzy lower and upper approximations of D_k can be computed more efficiently by considering only objects y belonging to the decision class D_k or objects with a fuzzy relation to objects in D_k exceeding a certain threshold [6, 10].

Definition 2.15 (Regions of the Extended Variable Precision Fuzzy Rough Set): For a fuzzy approximation space (U,R) and a fuzzy set $A \in F(U)$, with threshold $\beta \in [0,1]$.

- The β-positive region (certainly belonging region) of A is defined by:

$$POS_\beta A = \{x \in U : R_\beta A(x) > 1-\beta\}$$

- The β-boundary region (uncertain region) of A is defined by:

$$BND_\beta A = \{x \in U : 1-\beta \geq R_\beta A(x) > \beta\}$$

  or equivalently $BND_\beta A = \{x \in U : \overline{R}_\beta A(x) \geq 1-\beta \text{ and } R_\beta A(x) \leq 1-\beta\}$. The boundary region represents the overlap between the lower and upper approximations with an error rate β [10].

- The β-negative region (negation region) of A is defined by:

$$NEG_\beta A = \{x \in U : R_\beta A(x) \leq \beta\}$$

2.2.3. Proposed Dependency Measure

Based on the extended VPOFRS model, a dependency measure is developed for application in decision table data analysis.

Definition 2.16 (β-Positive Region of a Decision Table): For a decision table $NDT = (U,C,D)$, the β-positive region of a conditional attribute set $P \subseteq C$ with respect to the decision attribute set D is defined as the union of the β-positive regions of each decision class $D_k \in D$:

$$POS_\beta(P,D) = \bigcup_{D_k \in D} POS_\beta(R_P, D_k)$$

where $POS_\beta(R_P, D_k) = \{x \in U : R_{P\beta} D_k(x) > 1-\beta\}$ [10].

Definition 2.17 (β-Dependency of Decision Attributes on Conditional Attributes): The β-dependency of the decision attribute set D on the conditional attribute set $P \subseteq C$ is defined as the proportion of objects in the β-positive region of P with respect to D:

$$\gamma_\beta(P,D) = \frac{|POS_\beta(P,D)|}{|U|}$$

The value $\gamma_\beta(P,D) \in [0,1]$ reflects the consistency level of the decision table from the perspective of β-precision fuzzy rough sets [10, 11]. **Proposition 2.6** proves the monotonicity of this dependency measure: if $P \subseteq Q \subseteq C$, then $\gamma_\beta(P,D) \leq \gamma_\beta(Q,D)$ [11].

2.3. Proposed Attribute Reduction Method for Improving Classification Accuracy

Based on the proposed decision table consistency measure, this section presents an attribute reduction method to handle noise using the VPOFRS approach.

2.3.1. Proposed Model

Leveraging the proven monotonicity of the consistency measure γ_β, an attribute importance evaluation measure is constructed as follows:

Definition 2.18 (Attribute Importance): For a decision table $NDT = (U, C, D)$ and a subset of attributes $B \subseteq C$, the importance of an attribute $c \in C \setminus B$ with respect to B at level β is defined as the change in dependency when adding c to B:

$$Sig_\beta(c, B) = \gamma_\beta(B \cup \{c\}, D) - \gamma_\beta(B, D)$$

An attribute $c \in C \setminus B$ is considered unimportant with respect to B if $Sig_\beta(c, B) = 0$, meaning that adding c to B does not change the dependency [12].

Definition 2.20 (β-Reduct): For a decision table $NDT = (U, C, D)$, a subset $B \subseteq C$ is called a β-reduct if it satisfies the following conditions:

1. **Necessary Condition (Preservation)**: $\gamma_\beta(B, D) = \gamma_\beta(C, D)$. The reduced attribute set must preserve the consistency of the original attribute set.
2. **Sufficient Condition (Minimality)**: For every $b \in B, \gamma_\beta(B \setminus \{b\}, D) \neq \gamma_\beta(B, D)$. No attribute in the reduct is redundant.

[12, 13].

2.3.2. Proposed Algorithm

The VPOFRS_ARD (Variable Precision Optimized Fuzzy Rough Set based Attribute Reduction for improving classification accuracy) algorithm is proposed using the filter attribute approach with a strategy to find the reduct through the following steps [13, 14]:

1. Initialize the reduct set $B = \emptyset$.
2. Compute the initial dependency $\gamma_\beta(C, D)$.
3. Repeat until $\gamma_\beta(B, D) = \gamma_\beta(C, D)$:
   (a) For each attribute $c \in C \setminus B$, compute the importance $Sig_\beta(c, B) = \gamma_\beta(B \cup \{c\}, D) - \gamma_\beta(B, D)$.
   (b) Select the attribute $b \in C \setminus B$ with the highest importance. If multiple attributes have equal importance, choose one randomly.
   (c) Add the attribute b to the reduct set $B = B \cup \{b\}$.
4. Return the reduct set B.

Complexity analysis shows that the computational complexity of the VPOFRS_ARD algorithm is $O(|D||D_k||U||C|^2)$, where $|U|$ is the number of objects, $|C|$ is the number of conditional

attributes, $|D|$ is the number of decision classes, and $|D_k|$ is the number of objects in the k-th class [15]. Theoretically, since $|D_k| \leq |U|$, the VPOFRS_ARD algorithm executes faster than the traditional VPFRS algorithm [16].

2.3.3. Numerical Example Illustrating the Algorithm

This section presents a numerical example to illustrate the steps of the VPOFRS_ARD algorithm [16].

2.4. Experiments and Evaluation

This section presents experimental results to evaluate the effectiveness of the proposed algorithm in terms of attribute reduction time, classification accuracy, and the size of the obtained reduct [17].

2.4.1. Experimental Scenarios and Environment

The experimental scenarios include [17, 18]:

- Investigating the optimal β value range for each dataset to achieve the best balance between reduct size and classification accuracy.
- Comparing the VPOFRS_ARD algorithm with existing attribute reduction algorithms for improving accuracy, namely VPFRS [4] and IFD [19], based on three criteria: attribute reduction time, classification accuracy, and reduct size.

The experiments were conducted on datasets with low classification accuracy (indicating the presence of noise) [18] using the Python programming language on a Windows 10 platform with an i5 processor and 8GB RAM [20].

2.4.2. Evaluation of the Proposed Algorithm

This section evaluates the relationship between the β threshold, reduct size, and classification accuracy (using SVM and k-NN models) [20-28]. The results show that the reduct size tends to decrease as β increases, though there are differences across datasets [23, 24]. Notably, on the UFDC and SHDC datasets, classification accuracy does not decrease and even improves as β changes, demonstrating the algorithm's ability to effectively eliminate noisy attributes [25-28].

2.4.3. Comparison of Attribute Reduction Algorithms for Improving Classification Accuracy

2.4.3.1. Comparison of Reduct Sizes Obtained from the Algorithms

Figure 2.5 compares the reduct sizes among VPOFRS_ARD, VPFRS, and IFD [29-31]. IFD tends to produce the smallest reducts, but VPOFRS_ARD outperforms on the UFDC and UFDD datasets (the two with the lowest initial accuracy) [29, 30]. This indicates that VPOFRS_ARD is more effective at identifying and eliminating noisy attributes [30].

2.4.3.2. Comparison of Classification Accuracy of Reducts Obtained from the Algorithms

Figures 2.6 (SVM) and 2.7 (k-NN) compare the classification accuracy of the reducts [31-35]. Overall, the accuracy differences among the algorithms are minimal, but on the UFDC dataset, VPOFRS_ARD achieves significantly higher accuracy than IFD and VPFRS while reducing reduct size [32-35]. This further confirms VPOFRS_ARD's ability to handle noisy data.

2.4.3.3. Comparison of Execution Times of the Algorithms

Figure 2.8 compares the execution times of VPOFRS_ARD and VPFRS [35-38]. VPOFRS_ARD demonstrates superior time efficiency across all datasets, particularly on large datasets [36, 37]. This improvement is likely due to optimizations in the VPOFRS model, reducing the number of required computations [37].

2.5. Conclusion of Chapter 2

Chapter 2 has presented a new attribute reduction method based on extending the Variable Precision Fuzzy Rough Set (VPOFRS) model in fuzzy approximation spaces [38, 39]. The proposed method constructs an attribute importance measure based on dependency using the VPOFRS approach, enabling the selection of the most relevant attributes and the elimination of less relevant or noisy attributes [39].

Experimental results show that the VPOFRS_ARD algorithm executes faster than VPFRS and IFRS, producing smaller reducts with higher classification accuracy on certain noisy datasets, particularly UFDC [2, 38]. This confirms the effectiveness of the proposed method in mitigating the impact of noise and improving computational performance [2].

Beyond its application in attribute reduction, the VPOFRS model holds significant potential for broader applications in data analysis problems involving fuzzy and uncertain data [40].

CHAPTER 3. Incremental Computation Methods for Reduct Updates Using Information Granularity in Fuzzy Approximation Spaces

Chapter 3 of the dissertation focuses on addressing the problem of **incremental attribute selection** in dynamic data environments, specifically when the decision table undergoes changes in the object set [1]. This chapter stems from the limitations of traditional methods, which primarily rely on distance measures between partitions, in efficiently handling dynamic data. The dissertation proposes a new approach based on **extending the concept of knowledge information granularity to construct a measure in fuzzy approximation spaces** [1].

The main highlight of the chapter is the development of **incremental computation formulas** to efficiently update the reduced attribute set when there are changes in the object set (adding or removing objects) without recomputing from scratch [1]. The chapter also introduces **corresponding algorithms** for these updates, evaluated through experiments on real-world datasets [1, 2]. The results demonstrate that the proposed method is not only computationally efficient but also capable of improving the quality of the reduced attribute set [2].

3.1. Preliminary Knowledge

This section reviews some fundamental concepts regarding the **fuzzy equivalence relation matrix** of an attribute and the fuzzy equivalence relation matrix of an attribute set [3]. These relation matrices serve as the foundation for constructing computational tools in the subsequent sections of the chapter [4]. Details of these foundational concepts can be found in works [5], [6].

Definition 3.1 (Similarity Relation of Two Objects on an Attribute) [4]: For a decision table $NDT = \langle U, C, D \rangle$, the similarity relation of two objects $i, j \in U$ with respect to attribute $a \in C$ is defined by the formula:

$$sa_{ij} = 1 - |a(i) - a(j)| \quad (3.1) \tag{3.1}$$

where $|a(i) - a(j)|$ is typically normalized to the interval $[0, 1]$.

From the similarity relation of two objects on an attribute, the fuzzy equivalence relation matrix for each attribute and for an attribute set can be constructed by taking the intersection of the corresponding fuzzy equivalence relations [4].

3.2. Fuzzy Information Granularity Measure

This section introduces the extension of the information granularity measure from crisp approximation spaces to a measure in fuzzy approximation spaces, referred to as the **Fuzzy Information Granule (FIG) measure** [8]. First, the concept of the fuzzy information granule of an attribute is defined, followed by the construction of measures on fuzzy information granules, and finally, an attribute reduction algorithm based on the fuzzy information granularity measure for the decision table NDT is proposed [8].

3.2.1. Fuzzy Information Granule

Definition 3.4 (Fuzzy Information Granule) [8]: For a decision table $NDT = \langle U, C, D \rangle$ and an attribute set $P \subseteq C$, the fuzzy information granule of P on U is defined by:

$$[i]_P = \{(j, \mu_{[i]_P}(j)) | j \in U\} \tag{3.2}$$

where $\mu_{[i]_P}(j) = \mu_P(i, j)$ is the fuzzy similarity between objects i and j based on the attribute set P. The fuzzy similarity $\mu_P(i, j)$ is typically computed by taking the minimum (or another T-norm operator) of the similarities across each attribute in P:

$$\mu_P(i, j) = \min_{a \in P} \{sa_{ij}\} \tag{3.3}$$

The size of the fuzzy information granule $[i]_P$ can be defined as the sum of the membership degrees of all objects $j \in U$ to the granule $[i]_P$:

$$|[i]_P| = \sum_{j \in U} \mu_{[i]_P}(j) \tag{3.4}$$

3.2.2. Fuzzy Information Granularity Measure

Based on the concept of the fuzzy information granule, the **Fuzzy Information Granule (FIG) measure** of an attribute set P on U when considering the decision attribute D is defined

as follows [9]:

$$FIG(P \cup D, U) = 1 - \frac{1}{|U|^2} \sum_{i=1}^{|U|} |[i]_{P \cup D}| \tag{3.5}$$

This measure reflects the coarseness of the partition of the object set U based on the attribute set $P \cup D$. The smaller the value of $FIG(P \cup D, U)$, the finer the partition, and thus, the more informative it is.

Similarly, the fuzzy information granularity measure of P on U when considering only the conditional attribute set P is:

$$FIG(P, U) = 1 - \frac{1}{|U|^2} \sum_{i=1}^{|U|} |[i]_P| \tag{3.6}$$

The dependency of the decision attribute D on the conditional attribute set P based on the fuzzy information granularity measure is defined as:

$$FIG(D|P, U) = FIG(P, U) - FIG(P \cup D, U) = \frac{1}{|U|^2} \sum_{i=1}^{|U|} (|[i]_P| - |[i]_{P \cup D}|) \tag{3.7}$$

This measure indicates the extent to which knowing information from P reduces the coarseness of the partition when considering both P and D. The larger the value of $FIG(D|P, U)$, the higher the dependency.

Proposition 3.2 [9] proves the monotonicity of the fuzzy information granularity measure: If $P \subseteq Q \subseteq C$, then $FIG(P \cup D, U) \leq FIG(Q \cup D, U)$. This means that adding attributes to the conditional set reduces or maintains the coarseness of the partition (when considering both conditional and decision attributes).

3.2.3. Construction of the Attribute Reduction Algorithm

Based on the monotonicity of the fuzzy information granularity measure, **Definition 3.6** [9] defines the importance of an attribute $a \in C \setminus S$ with respect to an attribute set $S \subseteq C$ as:

$$Sig_U(a, S) = FIG(D|S \cup \{a\}, U) - FIG(D|S, U) \tag{3.8}$$

The attribute with the highest importance is prioritized for inclusion in the reduct.

Definition 3.7 [10] defines an attribute set $R \subseteq C$ as a reduct if it satisfies two conditions:

- $FIG(D|R, U) = FIG(D|C, U)$ (preservation of dependency).
- For every $a \in R$, $FIG(D|R \setminus \{a\}, U) < FIG(D|C, U)$ (minimality).

The **FIG (Fuzzy Information Granule)** algorithm [10] is proposed to find the reduct when the decision table NDT remains static. This algorithm is a **Greedy** algorithm using the **filter attribute** approach with an **attribute addition** strategy. The main steps of the algorithm include:

1. Initialize the reduct set $S = \emptyset$.
2. Compute $FIG(D|C, U)$ (dependency of the decision on the entire conditional attribute set).
3. Repeat until the stopping condition is met:
   (a) Compute the importance $Sig_U(a, S)$ for each attribute $a \in C \setminus S$.

(b) Select the attribute $s \in C \setminus S$ with the highest importance.

(c) Add s to the reduct set $S = S \cup \{s\}$.

(d) Check the stopping condition (e.g., $FIG(D|S,U) \approx FIG(D|C,U)$ or a desired number of attributes is reached).

4. Return the reduct set S.

The complexity of the FIG algorithm is analyzed as $O(|U|^2|C|^2)$ [11]. To optimize the reduct size, a threshold difference δ (e.g., 0.5%) between $FIG(D|S,U)$ and $FIG(D|C,U)$ can dissatisfied be used as the stopping condition [11].

3.3. Construction of the Reduct Update Algorithm for Adding Object Sets

This section presents the method for constructing an algorithm to update the reduct when the decision table NDT has an additional object set $\Delta u = \{u_1, u_2, ..., u_k\}$ [12]. The main steps include: developing an incremental computation formula for the fuzzy information granularity measure when adding a single object, extending it to the case of adding a set of objects, and finally proposing the corresponding reduct update algorithm (**FIGAMO - Fuzzy Information Granule based Attribute reduction for Adding Multiple Objects**) [12].

3.3.1. Incremental Formula for Adding a Single Object

Proposition 3.3 (Incremental Fuzzy Information Granule for Adding a Single Object) [12]: For a decision table $NDT = \langle U, C, D \rangle$ and a new object u_{new} added, forming the object set $U' = U \cup \{u_{new}\}$, the fuzzy information granule of the attribute set $P \subseteq C$ on U' is computed as follows:

$$FIG(P,U') = 1 - \frac{1}{|U'|^2}\left(\sum_{i\in U}|[i]_P| + |[u_{new}]_P| + \sum_{i\in U}\mu_P(i,u_{new}) + \sum_{i\in U}\mu_P(u_{new},i) + \mu_P(u_{new},u_{new})\right) \tag{3.9}$$

Similarly, the incremental formula for $FIG(P \cup D, U')$ is constructed [12]. From this, the incremental formula for the dependency $FIG(D|P,U')$ when adding a single object can be derived.

3.3.2. Incremental Formula for Adding a Set of Objects

Proposition 3.4 (Incremental Fuzzy Information Granule for Adding a Set of Objects) [13, 14]: For a decision table $NDT = \langle U, C, D \rangle$ and a set of objects to be added $\Delta u = \{u_1, u_2, ..., u_k\}$, the fuzzy information granule of P on $U \cup \Delta u$ is computed using an incremental formula based on $FIG(P,U)$ and the similarities between existing and new objects [14].

Proposition 3.5 (Incremental Fuzzy Information Granule Dependency for Adding a Set of Objects) [14, 15]: The incremental formula for $FIG(D|P, U \cup \Delta u)$ is constructed based on $FIG(D|P,U)$ and the similarities involving the new objects [15].

3.3.3. Proposed Incremental Algorithm for Adding a Set of Objects

Proposition 3.6 [15, 16] provides the formula for computing the importance of an attribute $a \in C \setminus B$ with respect to an attribute set B on $U \cup \Delta u$ based on the importance on U and the similarities involving the new objects [16].

The **FIGAMO** algorithm [16, 17] is proposed to update the reduct when adding the object set Δu. The main steps of the algorithm are:

1. Let RU be the reduct obtained on the decision table with the initial object set U. Initialize the new reduct set $S = RU$.
2. Compute $FIG(D|S, U \cup \Delta u)$ (Fs) and $FIG(D|C, U \cup \Delta u)$ (Fc) using the incremental formulas.
3. If the difference between Fs and Fc is less than a threshold ε (e.g., 0.5%), return S.
4. Otherwise, iterate by selecting the attribute with the highest importance (computed using the incremental formula) from $C \setminus S$ and adding it to S until the difference between $FIG(D|S, U \cup \Delta u)$ and $FIG(D|C, U \cup \Delta u)$ reaches the allowed threshold.
5. Remove redundant attributes from S by checking if removing an attribute significantly reduces the dependency.
6. Return the updated reduct set S.

The complexity of the FIGAMO algorithm is analyzed as approximately $O(|\Delta u|^2|C|)$ in the ideal case when the reduct does not change significantly [17].

3.4. Construction of the Reduct Update Algorithm for Removing Object Sets

This section presents the method for constructing an algorithm to update the reduct when the decision table NDT removes a set of objects Δu [18]. Similar to the addition case, the main steps include: developing an incremental computation formula for the fuzzy information granularity measure when removing a single object, extending it to the case of removing a set of objects, and finally proposing the corresponding reduct update algorithm (**FIGDMO - Fuzzy Information Granule based Attribute reduction for Deleting Multiple Objects**) [18].

3.4.1. Incremental Formula for Removing a Single Object

Proposition 3.7 (Incremental Fuzzy Information Granule for Removing a Single Object) [18]: For a decision table $NDT = \langle U, C, D \rangle$ and an object $u_{removed}$ removed, forming the object set $U' = U \setminus \{u_{removed}\}$, the fuzzy information granule of the attribute set $P \subseteq C$ on U' is computed by subtracting the contribution of $u_{removed}$ from the total [18]. Similarly, the incremental formulas for $FIG(P \cup D, U')$ and the dependency $FIG(D|P, U')$ when removing a single object are constructed.

3.4.2. Incremental Formula for Removing a Set of Objects

Proposition 3.8 (Incremental Fuzzy Information Granule for Removing a Set of Objects) [19]: For a decision table $NDT = \langle U, C, D \rangle$ and a set of objects to be removed $\Delta u = \{u_1, u_2, ..., u_k\}$, the fuzzy information granule of P on $U \setminus \Delta u$ is computed using an incremental formula based on $FIG(P, U)$ and the similarities involving the removed objects [20].

Proposition 3.9 (Incremental Fuzzy Information Granule Dependency for Removing a Set of Objects) [20]: The incremental formula for $FIG(D|P, U \setminus \Delta u)$ is constructed based on $FIG(D|P, U)$ and the similarities involving the removed objects.

3.4.3. Proposed Incremental Algorithm for Removing a Set of Objects

Proposition 3.10 [20] provides the formula for computing the importance of an attribute $a \in C \setminus B$ with respect to an attribute set B on $U \setminus \Delta u$ based on the importance on U and the similarities involving the removed objects.

The **FIGDMO** algorithm [21] is proposed to update the reduct when removing the object set Δu. Similar to FIGAMO, this algorithm uses incremental formulas to recompute the dependency and attribute importance, thereby efficiently updating the reduct [21].

3.5. Experiments and Evaluation of the Algorithms

This section presents experimental results to demonstrate the efficiency in terms of time and quality of the incremental attribute reduction algorithms FIGAMO and FIGDMO [22]. However, the experiments primarily focus on FIG and FIGAMO [22, 23].

3.5.1. Experimental Scenarios and Environment

The algorithms were implemented in Python on a Windows 10 platform with an i5 processor and 8GB RAM [24]. Standard datasets from the UCI Repository were used to evaluate the performance of the algorithms [24, 25]. The evaluation process included comparing execution time, reduct size, and classification accuracy (using SVM and k-NN models with 10-fold cross-validation) [24].

3.5.2. Evaluation of the FIG Algorithm

The FIG algorithm was compared with the FD (Fuzzy Dependency based attribute reduction) algorithm [6] on static datasets [24].

Execution Time

Figure 3.3 [26] shows that **the FIG algorithm has better execution time than the FD algorithm across all tested datasets** [26]. This demonstrates that the stopping condition (threshold difference in the information granularity measure) designed in the FIG algorithm helps reduce computation time [26].

Reduct Size

Figure 3.4 [27] shows that **the reduct size of the FIG algorithm is typically smaller or comparable to that of the FD algorithm** [27]. This confirms the FIG algorithm's ability to effectively reduce attributes [27]. A trade-off between execution time and reduct size is observed when varying the threshold difference [28].

Classification Accuracy

Figures 3.5 and 3.6 [28-31] compare the classification accuracy (using SVM and k-NN) of the reducts obtained from FIG and FD against the original attribute set. The results show:

- On small datasets, the classification accuracy shows no significant difference [29].

- On large datasets (CMSC, BCWD), **the reducts from FIG generally achieve higher classification accuracy than those from FD** [29, 30].

These results indicate that the FIG algorithm is effective in selecting important attributes and maintaining or improving classification performance [31].

3.5.3. Evaluation of the FIGAMO Algorithm

The FIGAMO algorithm was compared with the FDMAO (incremental FD) algorithm [6] when adding object sets [32].

Execution Time

Figure 3.7 [33] shows that **the FIGAMO algorithm has better execution time than the FDMAO algorithm across all tested datasets when adding objects** [33]. The stopping condition in FIGAMO continues to play a critical role in improving time efficiency [33]. The difference in execution time across different addition levels is generally small, but FIGAMO demonstrates greater stability than FDMAO on large datasets [34].

Reduct Size

Figure 3.8 [35] shows that **the reduct size of the FIGAMO algorithm is typically smaller or comparable to that of the FDMAO algorithm after adding objects** [35]. This indicates that FIGAMO maintains effective data compression in dynamic data environments [36].

Classification Accuracy

Figures 3.9 and 3.10 [37-40] compare the classification accuracy (using SVM and k-NN) of the reducts obtained from FIGAMO and FDMAO against the original attribute set after adding objects. The results show:

- On many datasets, **FIGAMO maintains classification accuracy equivalent to or higher than the original attribute set and often outperforms FDMAO** [38, 39].
- However, the effectiveness of attribute reduction may depend on the specific dataset [40, 41].

Overall, FIGAMO demonstrates the ability to maintain or improve classification accuracy after attribute reduction in environments with object set changes [41].

3.6. Conclusion of Chapter 3

Chapter 3 of the dissertation has presented an **efficient incremental attribute reduction method** using the information granularity approach in fuzzy approximation spaces [41, 42]. This method is designed to handle large datasets efficiently, particularly when data is continuously updated or changed [42].

The highlight of the method is the **extension of the information granularity measure to fuzzy approximation spaces**, which simplifies incremental computation formulas and significantly reduces the time required to update the reduct [42]. Experimental results have demon-

strated that the proposed algorithms (**FIG and FIGAMO**) not only have more efficient computation times compared to other methods but also produce reducts with smaller sizes and improved classification accuracy on some large datasets [43].

However, the dissertation also points out that the **selection of the adjustment threshold for the difference** is critical to balancing incremental computation time and the classification accuracy of the final reduct [43, 44]. This balance needs careful consideration in practical applications [44].

CONCLUSION

A. Main Results of the Dissertation

The dissertation contributes to the research stream on data preprocessing, applied to data mining and machine learning fields, focusing on attribute selection problems for numerical decision tables. Initially, the dissertation surveys the research gaps of existing methods and presents several key contributions as follows:

First, the dissertation proposes an *attribute selection method for decision tables based on an extended variable precision fuzzy rough set approach in fuzzy approximation spaces*.

The core of the method lies in constructing a measure to evaluate the importance of each attribute. This measure is based on dependency, defined using the VPOFRS approach, which enables precise determination of each attribute's impact on the data classification process. This facilitates the selection of the most important attributes while eliminating those that are less relevant or noisy.

To validate the effectiveness of the proposed method, the dissertation conducted experiments on various datasets. The results demonstrate that the new attribute reduction algorithm not only significantly improves computational time but also produces reducts with higher classification accuracy and smaller sizes compared to traditional algorithms. This indicates that the proposed method is not only faster but also more effective in simplifying models and enhancing their generalization capability.

Beyond its application to attribute reduction, the dissertation also highlights that the VPOFRS model has significant potential for broader applications in other data analysis problems. With its ability to handle fuzzy and uncertain data, VPOFRS can serve as a powerful tool in various domains, from pattern recognition to forecasting and decision-making, opening up numerous research and application avenues in the future.

Second, the dissertation proposes an *incremental computation method for reduct updates using the information granularity approach in fuzzy approximation spaces*.

The highlight of this method is the extension of the information granularity measure to fuzzy approximation spaces. By leveraging the concept of information granules, the computation formulas are simplified, significantly reducing the time required for incremental computations and reduct updates. This is particularly crucial when processing large datasets, where computational time becomes a critical factor.

Although the proposed algorithm shows promise in attribute reduction for large datasets, the research also indicates the need to investigate and adjust the threshold for difference adjustment before application. This is to ensure a balance between the computational time for incremental reduct updates and the classification accuracy of the final reduct. This balance is critical to achieving optimal performance in practical applications.

B. Future Directions of the Dissertation

The current research results of the dissertation are built upon fuzzy approximation spaces,

where numerical decision tables serve as input data. However, a limitation of this approach is that, in practice, many data analysis problems involve more complex decision tables that include not only numerical (continuous) attributes but also categorical attributes interspersed. The presence of both types of attributes in the same decision table poses unique challenges in the attribute reduction process.

Therefore, a potential direction for further development of the dissertation's research results is to extend the approach to neighborhood spaces. This extension would enable the application of the developed attribute reduction methods to mixed decision tables, encompassing both numerical and categorical attributes.

By extending to neighborhood spaces, the dissertation can address a broader range of real-world problems. This will enhance the applicability and practical value of the research results, as real-world data is often heterogeneous and contains various types of attributes.

LIST OF RESEARCH WORKS

[CT1] **Phạm Minh Ngọc Hà**, Nguyễn Long Giang, Nguyễn Văn Thiện, Nguyễn Bá Quảng, "Về một thuật toán gia tăng tìm tập rút gọn của bảng quyết định không đầy đủ", *Chuyên san các công trình nghiên cứu phát triển CNTT&TT, Tạp chí Công nghệ thông tin và truyền thông - Bộ TT&TT*, Tập 2019, Số 1, Tháng 9, Tr. 11-18.

[CT2] **Phạm Minh Ngọc Hà**, Nguyễn Long Giang, Trần Thanh Đại, Trần Văn Sinh, "Rút gọn thuộc tính trong bảng quyết định đầy đủ theo tiếp cận xác suất phân lớp của hạt thông tin lân cận mờ", *Hội thảo quốc gia lần thứ XXV Một số vấn đề chọn lọc của Công nghệ thông tin và truyền thông*, Hà Nội, 2022.

[CT3] **Pham Minh Ngoc Ha**, Tran Thanh Dai, Nguyen Manh Hung, Hoang Tuan Dung, "A Novel Extension Method of VPFRS Mode for Attribute Reduction Problem in Numerial Decision Tables," *J. Comput. Sci. Cybern.*, no.1 Mar, 2024.

[CT4] **Pham Minh Ngoc Ha**, Nguyen Long Giang, Nguyen Manh Hung, Tran Thanh Dai, "Incremental Attribute Reduction in Rough Set for Fuzzy Decision Tables", *Vietnam Journal of Science and Technology*, 2024.