MINISTRY OF EDUCATION
AND TRAINING

VIETNAM ACADEMY OF SIENCE
AND TECHNOLOGY

GRADUATE UNIVERSITY OF SIENCE AND TECHNOLOGY

Tran Thanh Dai

INTUITIONISTIC FUZZY ROUG SET AND TOPOLOGY BASED ATTRIBUTE REDUCTION IN DECISION TABLE

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

Supervisor 1: Associate Professor . Dr. Nguyen Long Giang, VAST, Ha Noi, Viet Nam

Supervisor 2: Professor . Dr. Vu Duc Thi, ITI, VNU, Ha Noi, Viet Nam

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……)

The dissertation can be found at:

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

INTRODUCTION

The urgency of the thesis topic

Attribute reduction [1]–[3] or attribute selection is an essential data preprocessing step, widely applied in fields related to pattern recognition and data mining, including data classification [4], [5], handwriting recognition [6], [7], speech recognition [8], [9], spam detection and classification [10], [11] and decision support [12], [13]. Attribute reduction aims to identify and select the subset of the original attribute set that is most relevant or eliminate redundant attributes that are least relevant to the decision-making of the problem. Attribute reduction is often performed so that the model achieves several goals, including increasing the understandability of the rules, improving performance, and reducing computational costs.

The Rough Set (RS) theoretical model introduced by Pawlack in 1982 is a powerful and effective mathematical tool for uncertain, incomplete, and inconsistent data [14]. Attribute reduction is one of the critical applications of rough set theory models, which has received the attention of researchers [15]–[17]. Based on the concept of equivalence class and approximation operations in the rough set theory model, many measure attribute importance are proposed to find the reduced set. Besides, topological space is also an essential concept in the rough set theory model [18], [19]. According to the rough set approach, the topological concept was also introduced by Pawlack in 1988 and received much attention from researchers [4], [20].

Over the past three decades, the direction of attribute reduction using the rough set approach [14] has been attracting the attention of many researchers. The research results show that the rough set approach to attribute reduction is effective on decision tables with discrete value attributes. However, with decision tables with continuous value attributes (numeric decision tables), it is necessary to convert the continuous value domain to the discrete value domain before reducing the attribute. This transformation step incurs implementation costs and may cause data loss during the transformation process. Therefore, the researchers propose reducing attributes directly on the original decision tables without going through the data discretization process.

Recently, Researchers have extended the traditional rough set theory model based on fuzzy sets (Fuzzy Set - FS) and intuitive fuzzy sets (Intuitionistic Fuzzy Set - IFS) to reduce the attributes directly on the original decision table, include:

1. Fuzzy Rough Set (FRS)

The fuzzy rough set model [21], [22] uses a similar concept instead of the indistinguishable concept in the traditional rough set theory model. Therefore, we do not need to discretize the data but still accurately evaluate the relationships of objects in a set. Up to now, research directions on attribute reduction using the fuzzy rough set approach have been quite exciting, with new proposals for measures including fuzzy positive domain measure (Fuzzy POS - FPOS) [17], [23]–[29], Fuzzy Information Entropy - FIE [13], [30]–[32], Fuzzy Distance - FS [33].

2. Intuitionistic Fuzzy Rough Set - IFRS

According to the definition of IFRS, each element in an intuitive fuzzy set is represented by two components: membership function and non-membership function. Evaluating each relationship of two objects based on these two components is said to be more rigorous than traditional fuzzy sets [34], [35]. Therefore, the algorithms according to the IFRS approach can improve classification accuracy for reduced sets better than the FRS approach in the case of noise data sets. In which noise data sets are data sets with low classification accuracy. Recently, some attribute reduction methods according to the IFRS approach include attribute reduction method according to the intuitionistic fuzzy positive domain approach (IF-POS) [36], according to the standard entropy approach Intuitionistic Fuzzy Information Entropy (IFIE) [15].

In Vietnam, several doctoral dissertations have been researching the method of reducing attributes directly on digital decision tables, including a doctoral dissertation by author Cao Chinh Nghia [3] researching the reduction of attributes on digital decision tables. Using fuzzy positive domain measures, calculate and generate decision rules on numerical data tables with fully defined domains. The doctoral thesis of author Nguyen Van Thien [2] proposes a fuzzy distance measure and builds some algorithms for finding the reduct according to the filter and wrapper methods. Author Ho Thi Phuong's doctoral thesis [1] proposes incremental algorithms for finding reduced sets in dynamic decision tables using fuzzy distance measures.

From the survey results above, direct attribute reduction methods on digital decision tables in Vietnam today are only based on the FRS approach. Experimental results show that the reduced set obtained by the FRS approach is ineffective in terms of size and classification accuracy on noisy data sets because the fuzzy approximation space is not enough to describe the relationship. of objects in a set. Regarding the IFRS approach [15], [36], the attribute reduction method in the world today still needs to be improved in terms of the size of the reduct and the execution time of the algorithm. The IF approximation proposed by the authors does not fully reflect an object's relational information, and the measurement of attribute significance still needs to be improved. Therefore, *the first research goal* of the thesis is to build an attribute reduction method based on the IFRS approach that is efficient in time and size and improves classification accuracy for noisy data sets.

Besides the attribute reduction methods following the rough set and extended rough set approaches as presented above. The topological method of attribute reduction also has received attention. Researchers have proposed it in recent years because topology's operational properties are similar to the RS model [37], [38].

According to the topological approach, the first concept of topology reduct introduced by Lashin and colleagues [37]. To reduce the attributes of the decision table according to the topological approach, it is first necessary to devise methods to make a topological structure based on the information already in the decision table, which is a big challenge, attracting the attention of many researchers [37]–[39]. Currently, there are two methods of building topology using the rough set approach, including methods of building topology from the approximate space of the rough set [38], [40]–[42], methods of building topology from rough set approximations [43]. Besides, the relationship of topological theoretical models and rough sets also attracts the attention of researchers [38], [43]–[47]. In particular, studies on the similarities between the approximation operations of the rough set theory model and the domain operations of the topology theory model [48]. On that basis, many topological structures are proposed based on the reconstruction of approximate operations of rough set theory [20], [45], [49]. Furthermore, based on this relationship, some rough set model reconstruction methods based on topological structure are also proposed [44], [50], [51].

However, most of the studies presented above are only theoretical general studies and approaches to building topology from rough sets and rough sets from topology to emphasize the close theoretical relationship between these two models. Recently, Xie et al. [52] proposed an attribute reduction method using the topological discriminant matrix approach. However, the research results still need to be improved regarding theoretical framework and applicability in practical data sets. Therefore, *the second research goal of the thesis* is to study the method of reducing attributes for decision tables according to the algebraic topology approach to construct a theoretical foundation for algebraic topology and apply it to attribute reduction.

Objectives of the study

From the remaining problems of current attribute reduction methods, the thesis sets out two research objectives, including 1) Research on attribute reduction methods according to the *IFRS*; 2) Research the method of reducing attributes according to the *algebraic topology* approach.

- *First research objective*: With the method of attribute reduction using the IFRS approach, *the first study* is to find out how to define the relationship of a practical object based on IFS, specifically building membership and non-membership evaluation functions for the IF approximate space. On that basis, *the subsequent study* is to construct a measure to evaluate the significance of attributes affected in terms of time and apply it to build an attribute reduction algorithm on noisy and high-dimensional data sets in practice.

- *Second research objective*: With the attribute reduction method using the belt topology approach, *the first study* is to learn methods of building topological structures, find out The basic properties of the topology are such that the topology can be evaluated in a smaller space to save computational costs. On that basis, *the subsequent study* is to study basic mathematical operations on topological structures to build methods to evaluate and identify the importance of attributes and define the reduct and apply to build an effective attribute reduction algorithm on high-dimensional data sets in practice.

Research subjects

The thesis focuses on researching attribute reduction methods on full decision tables with numerical value domains and noisy decision tables with a medium to large number of samples and dimensions. The thesis focuses on researching methods to reduce attributes in decision tables according to the rough set approach and algebraic topology, including:

- Survey the basic concepts of rough sets, measures used to evaluate the importance of attributes, and methods for building attribute reduction algorithms according to the Heuristic approach.

- Survey the basic concepts of topology according to the rough set approach, topology obtained from approximate space, topology obtained from the relationship of approximation operations, separability in topological space, and reduced topology.

Research scope

The thesis focuses on researching variations based on approaches of rough sets and algebraic topology based on FS and IFS, including:

- Research extended rough set models based on FS and IFS, applying them to build attribute reduction algorithms in numeric decision tables.

- Research topological structure according to rough set approach and some separability properties of topological space based on FS and IFS, apply to build attribute reduction algorithm in the numeric decision table.

Research methods:

The research results of the thesis are evaluated from two research perspectives including:

- *Theoretical research perspective*: Definitions and propositions are presented based on the basic foundation of set theory, measures, RS, FS, and IFS. - *Empirical research perspective*: algorithms are installed and tested on data sets from UCI[1]. Use data classification models suitable for the data and evaluation measures and methods to evaluate the quality of the reduced set. Comparing the quality of the reduct from the proposed algorithm with other algorithms to highlight the thesis's research hypothesis is completely reasonable.

Structure of the thesis:

Besides the introduction and conclusion, the thesis has 04 chapters with research content as follows:

Chapter 1. The thesis presents basic concepts and related research on the problem of attribute reduction according to RS and topological approaches. The main contributions of the thesis are presented in chapters 2, 3, and 4.

Chapter 2. The thesis presents an attribute reduction method based on the intuitive fuzzy rough set approach.

Chapter 3. The thesis presents an attribute reduction method based on an intuitive fuzzy topology approach.

Chapter 4. The thesis presents the attribute reduction method according to the Hausdorff topological approach.

Finally, the conclusion states the results achieved by the thesis, future development directions, and issues of concern of the author.

CHAPTER 1. OVERVIEW OF ROUGH SET AND TOPOLOGY APPROACH BASED ATTRIBUTE REDUCTION

1.1. Introduction

Attribute reduction, or feature selection, is a critical data preprocessing step in pattern recognition, machine learning, and data mining. For data sets for unsupervised learning problems, attribute reduction aims to select a subset of the original attribute set that preserves the information of the original attribute set. For data sets for supervised learning problems, attribute reduction aims to select a subset of the original attribute set that preserves the ability to classify or predict compared to the original set.

1.2. The basic concepts

1.2.1. Classical Rough Set

Definition 1.1 (Information System [14]). An information system is a quartet $IS = (U, A, V, f)$ where U is a finite nonempty set of objects, A is a finite nonempty set attributes, $V = \bigcup\limits_{a \in A} V_a$ where V_a is the set of values of attribute $a \in A$ and $f : U \times A \to V_a$ is the information function, $\forall a \in A,\ u \in U$ we have $f(u,a) \in V_a$.

Definition 1.2 (Attribute Partition [18], [53]). Given decision table $DT = (U, C, D, f)$ and $P, Q \subseteq C$. Then:

1) Partition U/P and partition U/Q are said to be the same or $U/P = U/Q$, if and only if $\forall u \in U,\ [u]_P = [u]_Q$.

2) A partition U/P is said to be finer than a partition U/P or $U/P \preceq U/Q$ if and only if $\forall u \in U,\ [u]_P \subseteq [u]_Q$

Definition 1.3 (Classical RS model [14], [18], [53]). In the traditional rough set theory

[1] https://archive.ics.uci.edu/ml/datasets.html

model, to represent the set $X \subseteq U$ on the knowledge base of the attribute set B according to the rough set concept, Pawlack uses two operations based on equivalence classes. of U/B. These operations are called B-lower approximation and B-upper approximation of X on U/B, denoted by $\underline{B}(X)$ and $\overline{B}(X)$. In there:

$$\underline{B}(X) = \{u \in U \,|\, [u]_B \subseteq X\} \tag{1.1}$$

$$\overline{B}(X) = \{u \in U \,|\, [u]_B \cap X \neq \emptyset\} \tag{1.2}$$

1.2.2. Intuitionistic Fuzzy Rough Set

Definition 1.4 (Intuitionistic Fuzzy Set [54])**.** Let U be a non-empty set of objects, the IFS X on U is determined by:

$$X = \{\langle x, \mu_X(x), \nu_X(x)\rangle \,|x \in U\} \tag{1.3}$$

In which, $\mu_X(x) \in [0,1]$ is the degree of membership of $x \in U$ with X and $\nu_A(x) \in [0.1]$ is the degree of non-membership of $x \in U$ with X such that $0 \leq \mu_X(x) + \nu_X(x) \leq 1 \forall x \in U$.

Then, for each traditional fuzzy set Y, the IFS X can be determined by:

$$X = \{\langle x, \mu_Y(x), 1 - \mu_Y(x)\rangle \,|x \in U\} \tag{1.4}$$

If $0 \leq \mu_X(x) + \nu_X(x) < 1$ then $\pi_X(x) = 1 - \mu_X(x) - \nu_X(x)$ is called the membership indecision of $x \in U$ with X.

Definition 1.5 (IFRS model [36])**.** Given the decision table $DT = (U,C,D,f)$, R is the fuzzy equivalence relation defined on U and $A \subseteq U$, we have:

$$\underline{A}(x) = \bigwedge_{y \in U} I(R(x,y), A(y)) \tag{1.5}$$

$$\bar{A}(x) = \bigvee_{y \in U} T(R(x,y), A(y)) \tag{1.6}$$

1.2.3. Topology space

The topological space [37] is denoted by the pair (U,τ), where U is a non-empty set of objects and τ is a family of subsets of U satisfying the following conditions:

(T1) $\Phi \in \tau$ and $U \in \tau$.

(T2) τ is closed under any union operation.

(T3) τ is closed under the finite intersection operation.

The pair (U,τ) is called a topological space defined on U with elements that are open sets and are subsets of U, the complements of open sets are called closed sets. .

Definition 1.6 (Base [55])**.** Let U be a non-empty set of objects. Then the base of the topology τ on U is a family of subsets of C denoted B so that:

(1) For each $x \in U$, there exists $G \subseteq U$ such that $x \in G$.

(2) For all $G_1, G_2 \in B$, if $x \in G_1 \cap G_2$, then there exists $G_3 \in B$ such that $x \in G_3$.

Definition 1.7 (Subbase [55])**.** Given the topological space (U,τ). Then $S \subseteq \tau$ is called a subbase of topology τ if the finite intersection of subsets of S forms the basis B of topology τ

Definition 1.8 (Tôpô Hausdorff [37])**.** Given an approximate space (U,τ), the topology $\tau_H \in (U,\tau)$ is called Hausdorff topology if every $x \neq y \in (U,\tau)$ always exists at two open neighbors $V_x, V_y \in \tau_H$ such that $V_x \cap V_y = \emptyset$.

1.2.4. The reduct

In the decision table, condition attributes are divided into three groups: core attributes, reductive attributes, and redundant attributes. Core attributes are indispensable attributes in accurately classifying a data set. The core attribute appears in all reduced sets of the decision table. Redundant attributes are attributes whose removal does not affect the classification of the data set; redundant attributes do not appear in any reduced set of the decision table. A reduced attribute is an attribute that appears in a specific reduced set of the decision table.

1.3. Some formulas for calculating membership degrees

1.3.1. Standardized data

(1) Min-max normalization:

$$F\left(f_{c_k}(x_i)\right) = \frac{f_{c_k}(x_i) - min_{c_k}}{max_{c_k} - min_{c_k}}\left(max_{c_k}^{'} - min_{c_k}^{'}\right) + min_{c_k}^{'} \tag{1.7}$$

Where max_{c_k} and min_{c_k} are the minimum and maximum values of the attribute c_k. After normalization, the attribute values are returned to the new segment $[min_{c_k}^{'}, max_{c_k}^{'}]$

(2) z-score normalization:

$$F\left(f_{c_k}(x_i)\right) = \frac{f_{c_k}(x_i) - \overline{c_k}}{\sigma_{c_k}} \tag{1.8}$$

In which, $\overline{c_k}$ and σ_{c_k} denote the average value and standard deviation of the attribute c_k.

1.3.2. Similarity measure

For discrete-valued attributes, the membership degree $r_{ij}^{c_{k_l}}$ is determined as follows:

$$r_{ij}^{c_{k_l}} = \begin{cases} 1, & \text{if } f_{c_{k_l}}(x_i) = f_{c_{k_l}}\left(x_j\right) \\ 0, & \text{otherwise.} \end{cases} \tag{1.9}$$

For attributes with numeric values, $r_{ij}^{c_{k_l}}$ can be determined by the function F as follows:

$$r_{ij}^{c_{k_l}} = F\left(x_i, x_j\right) \tag{1.10}$$

In which, F satisfies $F(x_i, x_i) = 1, F\left(x_i, x_j\right) = F\left(x_j, x_i\right)$, and $F\left(x_i, x_j\right) \in [0,1]$.

The following are some examples of function F

(1) $r_{ij}^{c_{k_l}} = 1 - \left|f_{c_{k_l}}(x_i) - f_{c_{k_l}}\left(x_j\right)\right|$.

(2) $r_{ij}^{c_{k_l}} = \max\left(\min\left(\frac{f_{c_{k_l}}\left(x_j\right) - f_{c_{k_l}}(x_i) + \sigma_{c_{k_l}}}{\sigma_{c_{k_l}}}, \frac{f_{c_{k_l}}(x_i) - f_{c_{k_l}}\left(x_j\right) + \sigma_{c_{k_l}}}{\sigma_{c_{k_l}}}\right), 0\right)$

In which, $\sigma_{c_{k_l}}$ is called the standard deviation.

1.4. Method for evaluating the reduct

1.4.1. Evaluation criteria

Current attribute reduction algorithms using the measured approach are often evaluated based on three criteria, including: *size* of the resulting reduced set, *classification accuracy* of the reduct on the specific model, and *execution time* of the algorithm.

The smaller the size of the reduced set obtained from the algorithm, the more efficient it is in model-building time. The higher the accuracy, the more influential the attribute selection method and the reduced set structure will be. The faster the execution time, the more influential the ability to reduce data on large data sets.

The general goal of attribute reduction algorithms is to achieve all three criteria above. However, in practice with noisy and complex data sets, the reduced set's size criteria and classification accuracy interest many researchers. The following are some metrics to evaluate the model's ability to classify accurately on the reduct.

1.4.2. Evaluation model and data

According to [56]'s survey, commonly used classification algorithms in evaluating the classification accuracy of data sets before and after reduction include decision tree model C .45, CART classification and regression trees, support vector machine SVM, and k-NN neighbor classifier model. For decision tables with numeric value domain attributes, k-NN and SVM classification models are used more than the remaining ones.

Most attribute reduction algorithms are researched and evaluated based on datasets downloaded from UCI. UCI dataset is a reliable, diverse database of topics, and many experts and researchers use it.

1.4.3. Evaluation index

(1) Accuracy:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}. \tag{1.11}$$

(2) Error:

$$Error = \frac{FP+FN}{TP+TN+FP+FN} \tag{1.12}$$

1.5. Some methods for shortening attributes

1.5.1. The discriminant matrix approach

In 1992, Skowron and Rauszer introduced the attribute reduction method based on the rough set basis [57]. Then the discriminant matrix has size $n \times n$ with $n = |U|$, denoted $M(DS) = \left(c_{ij}\right)_{n\times n}$ is identified by:

$$c_{ij} = \begin{cases} \left\{c \in C \mid c\left(x_i\right) \neq c\left(x_j\right)\right\}, \omega\left(x_i, x_j\right) \\ \emptyset, \text{ otherwise.} \end{cases} \tag{1.13}$$

In which: $\omega\left(x_i, x_j\right)$ satisfies one of the following conditions:
(1) $x_i \in \mathrm{POS}_C(D) \wedge x_j \notin \mathrm{POS}_C(D)$;
(2) $x_i \notin POS_C(D) \wedge x_j \in POS_C(D)$;
(3) $x_i, x_j \in POP_C(D) \wedge \left(x_i, x_j\right) \notin \operatorname{ind}(D)$.

The discriminant function of the discriminant matrix $f(C,D)$ is a Boolean function defined as follows:

$$f(C,D) = \wedge\left\{\vee c_{ij} \mid c_{ij} \neq \emptyset\right\} \tag{1.14}$$

Then the core attribute set is determined by:

$$\operatorname{core}_C(D) = \left\{c \mid c_{ij} = \{c\}\right\} \tag{1.15}$$

Up to now, there are quite a few attribute reduction methods based on the discriminant matrix approach proposed in works [58]–[61].

1.5.2. Attribute reduction method based on measure approach

1.5.2.1. Measure of dependence

The dependency measure introduced by [39] has received a lot of attention from researchers. The basis of this measure is based on the concept of positive domain (POS) of the rough set. Given a decision table $DT = (U, C, D, f)$ with $B \subseteq C$, $X \subseteq U$ and R as equivalence relations on U. Then the positive region of D according to B is determined as follows:

$$\mathrm{POS}_B(D) = \bigcup_{X_i \in U/D} \underline{R_B} X_i \tag{1.16}$$

Then, the dependence of D on B is determined by:

$$\gamma_B(D) = \frac{|P_B(D)|}{|U|} = \frac{\sum_{x \in U} P_B S_B(D)(x)}{|U|} \tag{1.17}$$

On that basis, the importance of the attribute according to the POS approach is determined based on the following two main formulas:

$$\mathrm{Sig}_1(a, B, D) = \gamma_B(D) - \gamma_{B-a}(D) \tag{1.18}$$

$$\mathrm{Sig}_2(a, B, D) = \gamma_{B \cup a}(D) - \gamma_B(D) \tag{1.19}$$

In which the formula 1.18 is suitable for the backward greedy search technique and the formula 1.19 is suitable for the forward greedy search technique.

On that basis, attribute reduction methods using the dependency approach are developed based on expanding these measures. Details of the methods are presented in Table 1.1.

1.5.2.2. Measure of certainty

Based on Shanon's concept of Information Entropy, three types of measures are expanded to evaluate information certainty including:

***Table 1.1:** Summary of methods to reduce attributes according to dependency*

ID	Reference	Data type	Approach	Background set	Evaluation standards
1	[62]–[73]	Hybrid	NRS	Classical	accuracy, size, compuation time
2	[27], [32], [74]–[79]	Number	NRS	FS	accuracy, size, compuation time
3	[80]	Number	NRS	IFS	accuracy, size, compuation time
4	[81]	Hybrid	PRS	Classical	accuracy, size, compuation time
5	[17], [22]–[29], [74], [76], [82]–[89]	Number	FRS	FS	accuracy, size, compuation time
6	[34]–[36], [80], [90]–[96]	Number	IFRS	FS	accuracy, size, compuation time

- Information entropy:

$$FE(B) = -\frac{1}{|U|}\sum_{i=1}^{|U|}\log_2\frac{\left|[x_i]_{R_B}\right|}{|U|} \tag{1.20}$$

- Combined entropy:

$$FE(B,E) = -\frac{1}{|U|}\sum_{i=1}^{|U|}\log_2\frac{\left|[x_i]_{R_B}\cap[x_i]_{R_E}\right|}{|U|} \tag{1.21}$$

- Conditional entropy:

$$FE(E \mid B) = -\frac{1}{|U|}\sum_{i=1}^{n}\log_2\frac{\left|[x_i]_{R_E}\cap[x_i]_{R_B}\right|}{\left|[x_i]_{R_B}\right|} \tag{1.22}$$

Then $\forall a \in C - B, B \subseteq C$, two methods to calculate the importance of attribute a with attribute set B are determined as follows:

$$\mathrm{Sig}(a,B) = FE(B) - FE(B - \{a\}) \tag{1.23}$$

$$\mathrm{Sig}(a,B,D) = FE(D \mid B - \{a\}) - FE(D \mid B) \tag{1.24}$$

On that basis, methods for reducing attributes according to the certainty approach are developed based on expanding these measures. Details of the methods are presented in Table 1.2.

1.5.2.3. Distance measure

Given decision table $DT = (U,C,D,f)$. For all $P,Q \subseteq C$, with the corresponding knowledge denoted by $K(P)$ and $K(Q)$. Where $K(P) = \{[u_i]_P\,|u_i \in U\}$ and $K(Q) = \left\{[u_i]_Q\,|u_i \in U\right\}$. Then, the knowledge gap between P and Q according to the Jacard approach is determined

Table 1.2: *Summary of methods to reduce attributes according to uncertainty*

ID	Reference	Data type	Approach	Background set	Evaluation standards
1	[15], [97], [98]	Number	Entropy thông tin	IFS	accuracy, size, compuation time
2	[31], [75], [99]	Number	Condition Entropy	FS	accuracy, size, compuation time
3	[100]	Hybrid	Combined Entropy	Classical	accuracy, size, compuation time
4	[101]	Number	Complement Entropy	FS	accuracy, size, compuation time

as follows:

$$d_J\left(K\left(P\right),K\left(Q\right)\right)=1-\frac{1}{|U|}\sum_{i=1}^{|U|}\frac{\left|\left[u_i\right]_P\cap\left[u_i\right]_Q\right|}{\left|\left[u_i\right]_P\cup\left[u_i\right]_Q\right|} \tag{1.25}$$

Then $\forall a\in C-B, B\subseteq C$, the importance of attribute a with attribute set B is determined as follows:

$$SIG_B\left(a\right)=d_J\left(K\left(B\right),K\left(B\cup D\right)\right)-d_J\left(K\left(B\cup\{a\}\right),K\left(B\cup\{a\}\cup D\right)\right) \tag{1.26}$$

On that basis, attribute reduction methods using the distance measure approach are developed based on expanding these measures. Details of the methods are presented in Table 1.3.

Table 1.3: *Summary of methods to reduce attributes by distance*

ID	Reference	Data type	Approach	Background set	Evaluation standards
1	[24], [33], [65], [102], [103]	Hybrid	KD	Classical, FS, IFS	accuracy, size, compuation time
2	[29], [104], [105]	Number	GD	FS	accuracy, size, compuation time
3	[29]	Number	PD	FRS	accuracy, size, compuation time

1.5.3. Attribute reduction method based on topological approach

Based on the concept of basis set β of topological space (U,τ). Lashin and colleagues [37] used the concept of redundancy relations to define the reductio set according to the topological approach as follows:

Definition 1.9 (The reduct according to topological approach [37]**).** Given a decision table $DT=(U,C,D,f)$, with $B\subseteq C$ and $r\in B$. Then r is called an unnecessary relation in B if: $\beta_B=\beta_{(B-\{r\})}$. Then: B is called a reduced set of C if and only if:

(i) $\beta_C=\beta_{(B)}$.

(ii) $\beta_C\neq\beta_{(B-\{\mathrm{r}\})}, \forall r\in C-B$.

Based on the definition of the reduced topology structure, a number of studies related to topology construction methods using the rough set approach are presented in Table 1.4

Table 1.4: *Summary of topology construction methods based on rough set approach*

ID	Reference	Basis of computation
1	[18], [20], [37], [39], [41], [106], [107]	Approximate space
2	[37]–[39], [41], [47], [48], [106]–[109]	Upper approximation set and lower approximation set
3	[20], [39], [45], [47], [55], [88], [108], [110], [111]	Sample space and relations of operations

1.6. Conclusion Chapter 1

Chapter 1 introduced an overview of the attribute reduction problem and classified attribute reduction methods. Presents important theoretical foundations for implementation in the next research chapters of the thesis.

CHAPTER 2. INTUITIONISTIC FUZZY ROUGH SET-BASED ATTRIBUTES METHOD IN DECISION TABLES

2.1. Introduction

In this chapter, the thesis presents the attribute reduction method based on the intuitive fuzzy distance measure approach. First, the thesis proposes a measure of the distance between two intuitive fuzzy partitions, based on which the thesis builds a measure to evaluate the importance of the attribute. Next, the thesis proposes a Heuristic algorithm to find reduced sets based on the proposed structure of reduced sets according to the δ - equal similarity approach. Finally, experiment and compare the proposed algorithm with the algorithms of A. Tan [36], [112] on data sets downloaded from UCI.

The results have been published in research works [CT3, CT4].

2.2. Constructing an IF distance measure

2.2.1. Distance between two IFS

Lemma 2.1 [IF numbers]**.** *Given three real numbers* $a, b, c \in [0,1]$. *Then:*

1) If $a \geq b$ *then* $a - b \geq \min(a,\ c) - \min(b,\ c)$

2) If $a \leq b$ *then* $a - b \leq max(a,\ c) - max(b,\ c)$

Proposition 2.1 (The relation of IFS)**.** *Let* $\overset{\approx}{X}, \overset{\approx}{Y}, \overset{\approx}{Z}$ *are IFS defined on U, where U is a non-empty set of objects. Then:*

1) If $\overset{\approx}{X} \subseteq \overset{\approx}{Y}$ *then* $\left|\overset{\approx}{Y}\right| - \left|\overset{\approx}{Y} \cap \overset{\approx}{Z}\right| \geq \left|\overset{\approx}{X}\right| - \left|\overset{\approx}{X} \cap \overset{\approx}{Z}\right|$

2) If $\overset{\approx}{X} \subseteq \overset{\approx}{Y}$ *then* $\left|\overset{\approx}{Z}\right| - \left|\overset{\approx}{Z} \cap \overset{\approx}{X}\right| \geq \left|\overset{\approx}{Z}\right| - \left|\overset{\approx}{Z} \cap \overset{\approx}{Y}\right|$

3) $\left|\overset{\approx}{X}\right| - \left|\overset{\approx}{X}\cap\overset{\approx}{Y}\right| + \left|\overset{\approx}{Z}\right| - \left|\overset{\approx}{Z}\cap\overset{\approx}{X}\right| \geq \left|\overset{\approx}{Z}\right| - \left|\overset{\approx}{Z}\cap\overset{\approx}{Y}\right|$

Proposition 2.2 (Distance betweent IFS)**.** *Let two IFS* $\overset{\approx}{X},\overset{\approx}{Y}$ *defined on U, with U is a non-empty set. Then* $\overset{\approx}{d}\left(\overset{\approx}{X},\overset{\approx}{Y}\right) = \left|\overset{\approx}{X}\cup\overset{\approx}{Y}\right| - \left|\overset{\approx}{X}\cap\overset{\approx}{Y}\right|$ *is a distance between two IFS* $\overset{\approx}{X},\overset{\approx}{Y}$*.*

2.2.2. Distance between two IF partitions

Definition 2.1 (Distance between two IF partitions)**.** Given the decision table $DT = (U,C,D,f)$ and $[\overset{\approx}{X}]$, $[X\overset{\approx}{\cup}D]$ respectively are partitions of X and $X\cup D$ with $X\subseteq C$. Then the distance between $[\overset{\approx}{X}]$, $[X\overset{\approx}{\cup}D]$ is determined by:

$$\overset{\approx}{d}\left([\overset{\approx}{X}],[X\overset{\approx}{\cup}D]\right) = \frac{1}{|U|^2}\sum_{i=1}^{|U|}\left(\left|[u_i]_{[\overset{\approx}{X}]}\right| - \left|[u_i]_{[\overset{\approx}{X}]}\cap[u_i]_{[\overset{\approx}{D}]}\right|\right) \tag{2.1}$$

Proposition 2.3 (Measure distance)**.** *Given the decision table* $DT = (U,C,D,f)$ *and* $[\overset{\approx}{X}]$, $[X\overset{\approx}{\cup}D]$ *respectively are partitions of* X *and* $X\cup D$ *with* $X\subseteq C$*. Then* $\overset{\approx}{d}\left([\overset{\approx}{X}],[X\overset{\approx}{\cup}D]\right)$ *is a measure distance.*

2.3. Attribute reduction in decision tables using IF measure distance

2.3.1. Hybrid filter-wrapper method, using IF distance measure

Definition 2.2 (Matrix δ equal)**.** Given two intuitive fuzzy relationship matrices $\overset{\approx}{M}_B = [b_{ij}]_{n\times n}$ and $M\ limits^{\approx}{}_C = [c_{ij}]_{n\times n}$ with $n = |U|$. Then $\overset{\approx}{M}_B$ and $\overset{\approx}{M}_C$ are said to be δ equal if and only if:

1) $\sup_{i,j=1}^{n}\left|\mu(b_{ij}) - \mu(c_{ij})\right| \leq 1-\delta$

2) $\sup_{i,j=1}^{n}\left|\nu(b_{ij}) - \nu(c_{ij})\right| \leq 1-\delta$

Where $\sup_{i,j=1}^{n}$ indicates the largest difference of two intuitive fuzzy relationship matrices achieved at position i,j, with $\delta\in[0.1]$. We denote $\overset{\approx}{M}_B \overset{\delta}{=} \overset{\approx}{M}_C$.

Definition 2.3 (Attribute significance)**.** Given decision table $DT = (U,C,D,f)$ and attribute set $B\subseteq C$. Then the importance of attribute $a\in C-B$ with attribute set B is determined by:

$$SIG_B(a) = \overset{\approx}{d}\left([\overset{\approx}{B}],[B\overset{\approx}{\cup}D]\right) - \overset{\approx}{d}\left([B\overset{\approx}{\cup}a],[B\cup\overset{\approx}{a}\cup D]\right) \tag{2.2}$$

Definition 2.4 (The reduct)**.** Given decision table $DT = (U,C,D,f)$ and attribute set $B\subseteq C$. Then the attribute set B is called a reduced set if:

1) $[B\overset{\approx}{\cup}D] \overset{\delta}{=} [C\overset{\approx}{\cup}D]$;

2) $\forall b\in B,\ \ [B-\{\overset{\approx}{b}\}\cup D] \overset{\delta}{\neq} [C\overset{\approx}{\cup}D]$.

Algorithms have complexity: $O\left(|C|\,|U|^2\right) + O(\mathbb{T}|\Delta||C|^2|U|^2) + O(\mathbb{T}|R_W^{\delta}|)$.

Algorithm 2.1 Two-stage filter-wrapper algorithm using intuitionistic fuzzy distance (IFD)

```
Input: $DT = (U, C, D, f)$, the classification model $Model$, $\Delta = \{0.1, 0.2, ..., 0.9\}$
Output: The reduct $R$
$R_W^A \leftarrow \emptyset$;
$R_W^\delta \leftarrow \emptyset$;
for all $c \in C$ do
    computation $\overset{\approx}{[c]}$;
end for
for all $\delta \in \Delta$ do
    $R_F^\delta \leftarrow \emptyset$;
    while $\overset{\approx}{[R_F^\delta \cup D]} \overset{\delta}{\neq} \overset{\approx}{[C \cup D]}$ do
        $c_m \in C - R_F^\delta | SIG_{R_F^\delta}(c_m) = \underset{c \in C - R_F^\delta}{Max} \left\{ SIG_{R_F^\delta}(c) \right\}$;        {Filter phase}
        $R_F^\delta := R_F^\delta \cup \{c_m\}$;
    end while
    if $ACC(Model, R_F^\delta) > ACC(Model, R_W^\delta)$ then
        $R_W^\delta = R_F^\delta$;        {Wrapper delta ($W_\delta$) phase}
    end if
end for
for $(i = 1; i < |R_W^\delta|; i++)$ do
    if $ACC(Model, R_W^\delta[0:i]) > ACC(Model, R_W^A)$ then
        $R_W^A = R_W^\delta[0:i]$;        {Wrapper attribute ($W_A$) phase}
    end if
end for
return $R_W^A$;
```

2.3.2. Experiment and evaluate the algorithm

Table 2.1: *Describe the reduct size obtained from the algorithms*

ID	Dataset	|C|	|R|			
			IFD-SVM	IFD-KNN	IFPOS[36]	IFIE[15]
1	heart	13	7	9	13	10
2	CMSC	20	11	11	20	20
3	PDS	22	9	7	8	10
4	BCWP	32	25	21	12	12
5	IS	34	16	5	11	19
6	UFDC	43	26	29	8	11
7	UFDD	43	27	25	6	8
8	SHDC	44	2	9	10	14
9	UFDB	51	2	2	5	11
10	DPDS	54	5	7	15	24
11	sonar	60	11	31	17	25
12	VRB	310	11	12	18	35

This chapter uses two algorithms by A. Tan [15], [36] to compare and evaluate the pro-

Table 2.2: *Compare the classification accuracy of the reducts on the SVM classification model*

ID	Dataset	|U|	Accuracy			
			Raw	IFD-SVM	IFPOS[36]	IFIE[15]
1	heart	270	84±0.7	84±0	84±0.6	82±0.7
2	CMSC	540	95±0.2	95±0.9	95±0.9	95±0.2
3	PDS	195	84±0.5	85±0.1	85±0.1	84±0.7
4	BCWP	198	77±0.2	77±0.1	76±0.7	76±0.5
5	IS	351	88±0	89±0.9	87±0.6	87±0.6
6	**UFDC**	181	44±0.1	52±0	49±1	49±0.3
7	UFDD	180	68±0.9	68±1	64±0.8	63±0.8
8	SHDC	267	79±0.6	79±0.5	79±0.8	79±0.9
9	UFDB	92	100.0	100.0	100.0	92±0.4
10	DPDS	170	98±0.5	98±0.5	98±0.7	98±0.3
11	**sonar**	208	65±0.3	70±0.5	70±0.2	64±0.7
12	VRB	126	83±0.7	88±0.7	91±0.2	80±0.5

Table 2.3: *Compare the classification accuracy of the reducts on the KNN classification model*

ID	Dataset	|U|	Accuracy			
			Raw	IFD-KNN	IFPOS[36]	IFIE[15]
1	heart	270	77±0.4	78±0.2	77±0.6	76±0.8
2	CMSC	540	84±0.9	86±0.9	84±0.4	84±0.6
3	PDS	195	85±0.5	87±0.8	87±0.3	84±0.3
4	BCWP	198	78±0.7	79±0.8	79±0.1	79±0.1
5	IS	351	85±0.3	92±0.5	88±0.6	88±0.6
6	**UFDC**	181	82±0.7	86±0.8	74±0.5	78±0.3
7	UFDD	180	81±0.8	84±0.2	77±0	82±0.1
8	**SHDC**	267	66±0.3	72±0.4	69±0.8	67±0.2
9	UFDB	92	99.0	100.0	100.0	98±0.8
10	DPDS	170	98±1	97±0.2	98±0.5	96±0.8
11	**sonar**	208	68±0.8	69±0.1	62±0.9	60±0.9
12	VRB	126	68±0.6	82±0.7	81±0.7	65±0.2

posed IFD algorithm. In which algorithm [36] is an algorithm that uses an intuitive fuzzy positive domain measure and algorithm [15] uses an Intuitionistic Fuzzy Information Entropy (IFIE) measure.

Table 2.1 compares the size of the reduced sets obtained from the algorithms. Tables 2.2 and 2.3 compare the classification accuracy of the reduced sets obtained from the algorithms on two SVM classification models and KNN.

2.4. Conclusion Chapter 2

Chapter 2 presented the attribute reduction method using the intuitive fuzzy rough set approach based on expanding the distance measure between partitions. Experimental results show that the proposed algorithm can improve classification accuracy on some noisy data sets.

CHAPTER 3. INTUITIONISTIC FUZZY TOPOLOGY-BASED ATTRIBUTE REDUCTION METHOD

3.1. Introduction

This Chapter proposes an attribute reduction method based on an intuitive fuzzy topology approach. First, we propose a topological structure based on an intuitive fuzzy priority relationship and, on that basis, research some properties of IF-base and IF-subbase to build a method to evaluate the similarity between two IF topologies. Next, we propose some attribute reduction algorithms based on the similarity of the two IF topologies and define the reduct according to the unit topology structure.

The research results in this Chapter are published in [CT2] and [CT6] awaiting round 2 review.

3.2. Proposing an IF topology structure

Definition 3.1 (IF relation)**.** Given a decision table $DT = (U,C,D,f)$, for all $(x,y) \in U$ and $\delta \in [0.5,\ rm1]$, Then $IFR_{\overline{a}}^{\geq}(x,y) = \langle y, \mu_y, \nu_y \rangle$ with $a \in C$ is determined by:

$$\mu_y = \begin{cases} 1 - |a(x) - a(y)|\ if\, p_a(x,y) \geq \delta \\ 0\, if\ other \end{cases} \qquad \nu_y = 1 - \mu_y \tag{3.1}$$

Where $p_a(x,y) = \frac{\mathrm{a}(x) - a(y) + 1}{2}$. Then, the value p_a always belongs to the interval $[0.5,1]$. When the value $\delta = 0.5$, this priority relationship is reflexive and transitive, when $\delta > 0.5$ this priority relationship is only transitive.

Definition 3.2 (IF-subbase)**.** Given decision table $DT = (U,C,D,f)$. Then the IF-subbase of $a \in C$ is defined by:

$$S_a = \left\{S_a^L, S_a^R\right\} \tag{3.2}$$

Where S_a^L and S_a^R are respectively the left IF-subbase corresponding to the relationship matrix $M_{\overline{a}}^{\geq}$ and the right IF-subbase corresponding to the relationship matrix system $\left(M_{\overline{a}}^{\geq}\right)^T$ on attribute $a \in C$, where $\left(M_{\overline{a}}^{\geq}\right)^T$ is the transition matrix predicate of matrix $M_{\overline{a}}^{\geq}$.

Definition 3.3 (Intersect two IF-subbases)**.** Given a decision table $DT = (U,C,D,f)$ and two IF-subbases S_p, S_q corresponding to $p,q \in C$. Then, the intersection operation of two IF-subbases is defined by:

$$S_p \cap S_q = \left\{S_p^L \cap S_q^L, S_p^R \cap S_q^R\right\} \tag{3.3}$$

Definition 3.4 (Union of two IF-subbases)**.** Given a decision table $DT = (U,C,D,f)$ and two IF-subbases S_p, S_q corresponding to $p,q \in C$. Then, the union operation of two IF-subbases is defined by:

$$S_p \cup S_q = \left\{S_p^L \cup S_q^L, S_p^R \cup S_q^R\right\} \tag{3.4}$$

Definition 3.5 (IF-base)**.** Given decision table $DT = (U,C,D,f)$ and IF-subbase $S_a = \left\{S_a^L, S_a^R\right\}$ corresponding to $a \in C$, where S_a^L is called the left IF-subbase and S_a^R is called the right IF-subbase. Then IF-base B_a is defined by:

$$B_a = {S_a}^L \cap {S_a}^R \tag{3.5}$$

Proposition 3.1 (IFT from IF-base)**.** *Let the decision table $DT = (U,C,D,f)$ and B_a be an IF-base determined by the formula 3.5. Then, B_a is a basis of $\mathscr{T}_a$.*

Definition 3.6 (IF-subbase of attributes)**.** Given a decision table $DT = (U,C,D,f)$, for all $p,q \in C$. Then the IF-subbase of $\{p\} \cup \{q\}$ is defined by:

$$S_{\{p\}\cup\{q\}} = S_p \cap S_q \tag{3.6}$$

Definition 3.7 (Smoothest IF-base)**.** Given a decision table $DT = (U,C,D,f)$ and an IF-base B_a equivalent to $a \in C$. Then B_a is called the smoothest IF-base if: $B_a[i,j] = \begin{cases} 1_{IF} \ if \ i = j \\ 0_{IF} \ if other \end{cases}$

Where $1_{IF} = (1,0)$ and $0_{IF} = (0,1)$. The smoothest IF-base notation is B_I which is the basis of intuitive fuzzy unit topology.

Algorithm 3.1 IFT-based attribute reduction using filter method (F_IFT)

Input: The decision table $DT = (U,C,D,f)$ và $\delta = 0.5$

Output: The reduct R

```
R ← ∅;
B_R is the coarsest IF-base;
B_I is the smoothest IF-base;
for all c ∈ C ∪ D do
    calculate S_c;                                    {by formula 3.1 và 3.2}
end for
while B_R ≠ B_I do
    for all c ∈ C − R do
        calculate Sig_R(c);                           {by formula 3.9}
    end for
    select c_m ∈ C − R : Sig_R(c_m) = Max_{c∈C−R} {Sig_R(c)};
    R ← R ∪ {c_m};
    update B_R;                                       {by formula 3.5}
end while
return R;
```

3.3. Similarity measure of two intuitive fuzzy topologies

Proposition 3.2 (Difference between two IF-subbases)**.** *Given a decision table $DT = (U,C,D,f)$ and two IF-subbases S_p, S_q corresponding to $p,q \in C$. Then:*

$$\zeta(S_p,S_q) = \frac{2}{|U|^2} \sum_{i=1}^{|U|} \left(\left|S_p^L[i] \cup S_q^L[i]\right| - \left|S_p^L[i] \cap S_q^L[i]\right| \right) \tag{3.7}$$

Is the difference between S_p and S_q

Proposition 3.3 (Dependency of the attribute according to IF-subbase)**.** *Given a decision table $DT = (U,C,D,f)$ and two IF-subbases S_C and $S_{C\cup D}$ corresponding to C and $C \cup D$.*

Then:

$$\zeta\left(S_C, S_{C\cup D}\right) = \frac{2}{|U|^2}\sum_{i=1}^{|U|}\left(\left|S_D^L[i] - S_D^L[i]\cap S_C^L[i]\right|\right) \tag{3.8}$$

Is the dependency of attribute D with attribute C.

Proposition 3.4 (Anti-monotonicity of similarity measure)**.** *Given a decision table $DT = (U, C, D, f)$ and two IF-subbases S_B, S_C corresponding to B and C. Then, if $B \subseteq C$ then $\zeta\left(S_D, S_{D\cup C}\right) \leq \zeta\left(S_D, S_{D\cup B}\right)$:*

Algorithm 3.2 Hybrid filter - wrapper attribute reduction method using IFT approach (FW_IFT)

Input: The decision table $DT = (U, C, D, f)$ và $\delta = 0.5$, the classification $Model$

Output: The reduct R

```
 1: $ST \leftarrow \emptyset$; $R_W \leftarrow \emptyset$; $R_F \leftarrow \emptyset$; $R \leftarrow \emptyset$;
 2: $B_{R_F}$ is the coarsest IF-base;
 3: $B_I$ is the smoothest IF-base;
 4: for all $c \in C \cup D$ do
 5:     calculate $S_c$;                                  {by formula 3.1 và 3.2}
 6: end for
 7: for all $c \in C - R_F$ do
 8:     calculate $Sig_{R_F}(c)$;                         {by formula 3.9}
 9: end for
10: for all $c_m \in \{\underset{c\in C-R_F}{Max}\{Sig_{R_F}(c)\}\}$ do
11:     $ST.PUSH(R_F \cup \{c_m\})$;                      {Push $c_m$ base on Stack}
12: end for
13: while $ST \neq \emptyset$ do
14:     $R_F = ST.POP$;                                   {Filter Phase}
15:     update $B_{R_F}$
16:     if $B_{R_F} = B_I$ then
17:         $R_W = R_W \cup \{R_F\}$;
18:     else
19:         go back step 10;
20:     end if
21: end while
22: for all $r \in R_W$ do
23:     if $ACC(Model, r) > ACC(Model, R)$ then
24:         $R = r$;                                      {Wrapper Phase}
25:     end if
26: end for
27: return $R$;
```

3.4. Attribute reduction in the decision table using IF topology approach

3.4.1. Propose an algorithm to find a reduct in the decision table using the filter method.

Definition 3.8 (Attribute significance)**.** Given decision table $DT = (U, C, D, f)$ and attribute set $R \subseteq C$. Then, the importance of attribute $a \in C - R$ with attribute set R is defined by:

$$Sig_R(a) = \zeta(S_D, S_{D \cup R \cup a}) - \zeta(S_D, S_{D \cup R}) \tag{3.9}$$

Proposition 3.5 (Existence of the reduct)**.** *Given a decision table $DT = (U, C, D, f)$ and two IF-bases B_R and B_C corresponding to $R \subseteq C$. Then, if $B_R = B_I$ then $B_C = B_I$.*

Based on the clause 3.5 we can confirm that if a decision table exists a subset R of the original attribute set C for which B_R is the smoothest basis then certainly B_C is also the smoothest basis. Meaning $B_R = B_C = B_I$. Then, the reduced set can be defined as follows:

Definition 3.9 (The reduct base on unit IFT)**.** Given decision table $DT = (U, C, D, f)$ and $R \subseteq C$. Then R is called a reduced set of C if and only if

(1) $B_R = B_I$

(2) $B_{R-c} \neq B_I$ for all $c \in R$

To ensure the existence of B_I, the proposed intuitive fuzzy priority relationship must have reflexive properties, so the default δ value is chosen to be 0.5 for all illustrative examples and experiments with algorithms.

The F_IFT algorithm has complexity: $\mathcal{O}\left(|R|\,|C-R|\,|U|^2\right)$ and the FW_IFT algorithm have complexity: $\mathcal{O}\left(|ST|\,|C-R_F|\,|U|^2\right) + \mathcal{O}\left(|R_{\mathrm{W}}|\,|T|\right)$.

3.4.2. Experiment and evaluate algorithms

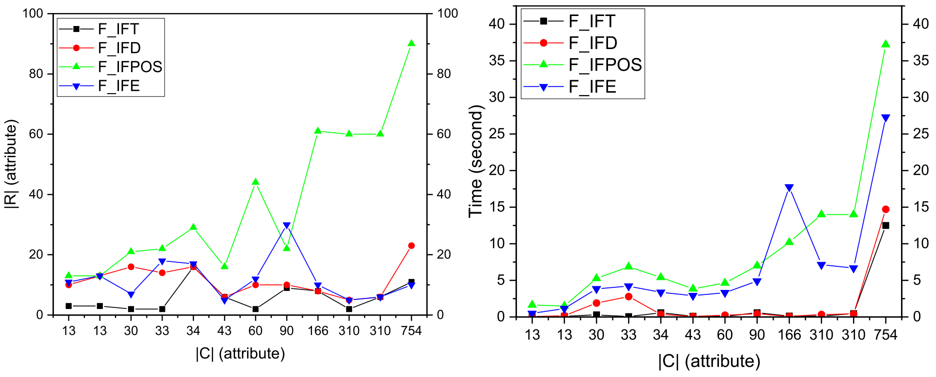

Figure 3.1: *Graph evaluating the relationship of reduce set size (left) and execution time (right) with the number of initial attributes of the F_IFT algorithm compared to other algorithms*

This section will present some experimental results of the two proposed algorithms. In which the F_IFT algorithm will be compared with the algorithms of A. Tan [15], [36] and Thang [102]. The FW_IFT algorithm will be compared with the FW_IFD algorithm [102].

Figure 3.1 shows the advantages in execution time of the F_IFT algorithm and the relationship between time and the resulting reduced set size of the algorithm. Meanwhile, Tables

3.1 and 3.2 show the reduct's size and classification accuracy advantages from the FW_IFT algorithm.

Table 3.1: *Compare the size of the reduced sets obtained from algorithms using the filter - wrapper approach on the SVM and KNN classification models*

ID	Dataset	FW_IFT		FW_IFD		$\|C\|$
		SVM	KNN	SVM	KNN	
1	Wine	5	4	10	10	13
2	Heart	6	5	11	11	13
3	Wdbc	3	5	16	16	30
4	Wpbc	3	2	2	2	33
5	Iono	7	5	12	12	34
6	UFDC	8	6	5	5	43
7	Sona	3	2	9	9	60
8	Libras	18	13	7	14	90
9	Musk	5	5	3	3	166
10	LVB	6	2	2	2	310
11	LVG	7	5	5	5	310
12	PD	9	11	17	23	754

Table 3.2: *Compare the classification accuracy of reduced sets obtained from algorithms using the filter - wrapper approach on the SVM and KNN classification models*

ID	Data	FW_IFT		FW_IFD		$\|C\|$	
		SVM	KNN	SVM	KNN	SVM	KNN
1	Wine	94.24	91.25	97.87	94.74	98.16	96.25
2	Heart	86.43	78.85	84.65	76.74	84.5	77.44
3	Wdbc	97.15	95.42	97.99	95.02	98.33	95.45
4	Wpbc	77.79	76.12	76.14	78.34	78.02	77.18
5	Iono	87.1	92.05	85.46	89.14	88.37	86.04
6	**UFDC**	68.16	90.9	50.95	69.14	43.49	79.13
7	**Sona**	77.21	68.35	67.35	61	65.45	68.16
8	Libras	70.9	77.59	64.79	78.02	71.41	75.23
9	Musk	73.17	75.13	62.51	64.41	75.54	77.37
10	LVB	85.29	77.19	77.71	76.31	83.24	67.8
11	LVG	90.22	78.64	70.18	66.93	89.05	69.22
12	PD	84.47	84.79	84.8	65.53	81.26	81.8

3.5. Conclusion Chapter 3

Chapter 3 presented the attribute reduction method using the IF topology approach and proposed two algorithms. The reduct of the F_IFT algorithm has an efficient size and execution time, while the FW_IFT algorithm for the reduced set has an effective size and classification accuracy.

CHAPTER 4. HAUSDORFF TOPOLOGY BASED ATTRIBUTE REDUCTION

4.1. Introduction

This section summarizes research results on the attribute reduction method according to the Hausdorff topological approach. In particular, propose the topological structure according to the rough set approach on the fuzzy approximation space β and propose the Hausdorff topological structure, proposing the concept of co-dependent structure in Hausdorff topological space, proposing an algorithm to find a reduced set based on a new definition of essential attributes according to the Hausdorff topology approach and grouping attributes according to the dependency structure concept of Hausdorff topology.

The research results in this Chapter are published in research works [CT1] and [CT5]

4.2. Proposing a topological structure from a thresholded fuzzy β

Definition 4.1 (The threshold fuzzy relationship formula β)**.** The threshold fuzzy equivalence relationship β of $u_i, u_j \in U$ is determined as follows:

$$R^{\beta}\left(u_i,u_j\right) = \{\ 1-\left|u_i-u_j\right| : if\ 1-\left|u_i-u_j\right| \geq \beta\ 0 : if\ 1-\left|u_i-u_j\right| < \beta. \tag{4.1}$$

Proposition 4.1 (Topology struct base on rough set)**.** *Let the approximation space* (U,R^{β}) *and* R^{β} *be a fuzzy equivalence relation. Then* $\mathscr{T} = \left\{X \subseteq U | \underline{R^{\beta}}(X) = \overline{R^{\beta}}(X)\right\}$ *is a topology defined on* U*.*

4.3. Proposed Hausdorff topology

Definition 4.2 (Separability of the fuzzy relation threshold β)**.** Let the approximation space (U,R^{β}) where R^{β} is the fuzzy equivalence relation β. Then R^{β} is said to be distinguishable if for every $u_i \in U$ there exists $u_j \neq u_i \in U$ such that $[u_i]_{R^{\beta}} \cap [u_j]_{R^{\beta}} = \emptyset$. The symbol for this relationship is R^{β}_{H}.

Proposition 4.2 (Hausdorff topology from relation R^{β}_{H})**.** *Let the topology* $\mathscr{T}_H = \{X \subseteq U | \underline{R^{\beta}}(X) = \overline{R^{\beta}}(X)\}$ *defined on* U*. Then,* $\mathscr{T}_H$ *is called a Hausdorff topology if* R^{β} *is an* R^{β}_{H}*.*

Proposition 4.3 (Determine the attribute with the relationship R^{β}_{H})**.** *Given decision table* $DT = (U,C,D,f)$ *and* $c \in C$*. Then* c *is called an attribute with relationship* R^{β}_{H} *if* $max_1(V_c) - max_2(V_c) > \beta$*. Where* V_c *is the set of values of attribute* c*.*

4.4. Attribute reduction in decision table according to Hausdorff topological approach

4.4.1. Proposing an algorithm to find the reduct in a decision table based on Hausdorff topology structure using the filter-wrapper hybrid method

Definition 4.3 (The signification of attribute according to the Hausdorff topological approach)**.** Given decision table $DT = (U, C, D, f)$ and $c \in C$. Then c is called an important property for D if $\mathscr{T}_c$ is a Hausdorff topology.

Definition 4.4 (Dependent co-structure)**.** Given a decision table $DT = (U, C, D, f)$ and two topologies $\mathscr{T}_p$, $\mathscr{T}_q$ defined on U corresponding to $p, q \in C$. Then $\mathscr{T}_p$ is said to be co-dependent with $\mathscr{T}_q$ if $\mathscr{T}_p \cup \mathscr{T}_D = \mathscr{T}_q \cup \mathscr{T}_D$.

Algorithm 4.1 Attribute reduction algorithm using filter - wrapper attribute clusters approach (CFW).

Input The decision table $DT = (U, C, D)$ with $\Delta = \{0.1, 0.2, ..., 0.8, 0.9\}$ and classification $Model$

Output The reduct R

```
R = ∅;
for all $\beta \in \Delta$ do
    $H^\beta \leftarrow \emptyset$;
    $CH^\beta \leftarrow \emptyset$;
    $R^\beta \leftarrow \emptyset$;
    for all $c \in C$ do
        if $max_1(V_c) - max_2(V_c) > \beta$ then
            $H^\beta = H^\beta \cup \{c\}$;                    {Filter Hausdorff attribute}
        end if
    end for
    for all $p \in \{H^\beta - CH^\beta\}$ do
        $U_p = \emptyset$;
        for all $q \in \{H^\beta - CH^\beta - p\}$ do
            if $\mathscr{T}_p \cup \mathscr{T}_D = \mathscr{T}_q \cup \mathscr{T}_D$ then
                $U_p = U_p \cup \{q\}$;                    {Clustering Hausdorff attribute}
            end if
        end for
        $CH^\beta = CH^\beta \cup U_p$;
        if $ACC_{U_p}^{Model}$>$ACC_{R^\beta}^{Model}$ then
            $R^\beta = U_p$;                    {Wrapper Hausdorff atribute group}
        end if
    end for
    if $ACC_{R^\beta}^{Model}$>$ACC_{R}^{Model}$ then
        $R = R^\beta$;
    end if
end for
return $R$;
```

4.4.2. Experiment and evaluate the algorithm

The proposed algorithm is compared and evaluated with typical attribute reduction algorithms based on the measure approach, including: (1) attribute reduction algorithm based on rough set approach with adjusted precision (VPRS).) [113]; (2) attribute reduction algorithm using fuzzy rough set (FRS) approach [114]; (3) attribute reduction algorithm according to Fuzzy Information Entropy (IFE) approach [82]; (4) attribute reduction algorithm using fuzzy distance (FD) approach [33].

Table 4.1 compares and evaluates the size of the reduced set obtained by the algorithms. In addition, the relationship between the size of the data set and the execution time of the algorithms is also described in Figure 4.1. Tables 4.2 and 4.3 compare the classification accuracy of the reduced set obtained from algorithms on two classification models k-NN and SVM.

Table 4.1: *Compare the size of the reduct set obtained from the algorithms*

ID	Dataset	$\lvert R \rvert$						
		$\lvert C \rvert$	CFW-SVM	CFW-kNN	VPRS	FRS	FIE	FD
1	wine	13	10.8	7.6	11.8	10.4	10.6	7.1
2	heart	13	6.7	5.5	11.5	13.9	10.2	6.7
3	CMSC	20	8.2	8.7	9.5	20.3	20.1	3.5
4	PDS	22	5.2	4.4	9.4	8.5	10.8	4.3
5	BCWD	30	3.2	3.6	14.8	7.6	12.1	4.1
6	BCWP	32	2.9	2.2	8.9	12.6	12.4	5.8
7	IS	34	2.1	2.1	20.9	11.3	19.6	6.1
8	UFDC	43	13.9	4.3	15.3	8.7	11.7	5.2
9	UFDD	43	5.1	3.6	19.9	6.6	8.3	3.3
10	SHDC	44	3.1	2.2	44.3	10.3	14.7	5.9
11	UFDB	51	4.1	3.4	8.9	5.8	11.9	5.2
12	DPDS	54	2.5	1.6	8.4	15.7	24.4	4.4
13	sonar	60	4.6	7.4	44.3	17.5	25.2	7.6
14	musk	166	5.7	11.4	86.6	23.9	29.5	8.8
15	VRB	310	9.1	4.3	56.6	18.9	35.8	7.5
16	VRG	310	9.6	2.1	72.4	16.5	36.4	10.6

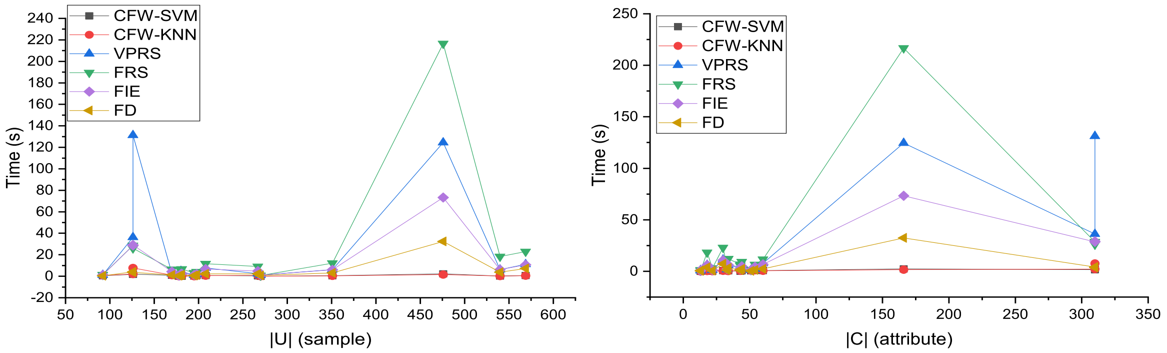

Figure 4.1: *Diagram analyzing the relationship between the algorithm's execution time and $|U|$ (left), between the algorithm's execution time and $|C|$ (right).*

Table 4.2: *Compare the classification accuracy of the reduced set obtained from algorithms on the SVM classification model*

ID	Dataset	Classification Accuracy (%)					
		Rawset	CFW-SVM	VPRS	FRS	FIE	FD
1	wine	98±0.7	96±0.9	99±0.6	99±0.3	93±0.1	96±0.8
2	heart	84±0.8	86±0.6	84±0.3	84±0.3	82±0.9	80±0.7
3	CMSC	95±0.8	95±0.4	92±0.4	95±0.1	95±0.8	92±0.6
4	PDS	84±0.7	86±0.6	84±0.7	85±0.9	84±0.7	75±0.8
5	BCWD	98±0.6	94±0.7	94±0.2	96±0	96±0.8	94±0.7
6	BCWP	77±0.3	76±0.3	76±0.6	76±0.2	76±0.8	76±0
7	IS	88±0.5	82±1	88±0.9	87±0.5	87±0.3	89±0.6
8	UFDC	44±0.8	**59±0.7**	45±0.5	49±0.1	49±0.6	50±1
9	UFDD	68±0.8	63±0.5	68±0.1	64±1	63±0.7	62±0.5
10	SHDC	79±0.5	79±1	79±0	79±0	79±0.6	79±0.3
11	UFDB	100±0.4	96±0.9	100±0.6	100±0.2	92±0.8	100±0.2
12	DPDS	98±0.6	98±0.3	98±0.3	98±0.6	98±0.4	98±0.5
13	sonar	65±0.8	**73±0.2**	65±0.2	70±0.7	64±0	58±0
14	musk	75±0.3	72±0.2	74±0.8	61±0.4	61±0.1	55±0.4
15	VRB	83±0.1	83±0.2	88±0.6	91±0.4	80±0.8	86±1
16	VRG	85±0.9	80±0.2	91±0.7	82±0.5	67±0.2	68±0.4

Table 4.3: *Compare the classification accuracy of the reduced set obtained from algorithms on the KNN classification model*

ID	Dataset	Classification Accuracy (%)					
		Rawset	CFW-kNN	VPRS	FRS	FIE	FD
1	wine	96±0.2	94±0.1	94±0.1	96±0.9	91±0.4	94±0.6
2	heart	77±0.5	78±0.1	77±0.3	77±0.3	76±0.2	69±0.7
3	CMSC	84±0.1	92±0.1	86±0.2	84±0.6	84±0.9	71±0.1
4	PDS	85±0.7	85±0.3	88±0.9	87±0.1	84±0.3	74±0.5
5	BCWD	95±0.2	93±0.1	93±0.9	93±0.9	94±0.7	93±0.7
6	BCWP	78±0.8	81±0.9	74±0.6	79±0.6	79±0.6	75±0.6
7	IS	85±0.6	88±0.6	86±0.9	88±0.7	88±0.4	89±0.4
8	UFDC	82±0.1	96±0.2	82±0.1	74±0.9	78±0.1	76±0.2
9	UFDD	81±0.5	81±0.9	77±0.9	77±0.5	82±0.6	72±0.7
10	SHDC	66±0.1	75±0.7	66±0.5	69±0.8	67±1	72±0.6
11	UFDB	99±0.8	100	100	100	98±0	99±0.5
12	DPDS	98±0.4	98±0	98±0.3	98±0.4	96±0.9	98±0.2
13	sonar	68±0.3	71±0.3	64±0.5	62±0.7	60±0.7	55±0.3
14	musk	77±0.5	76±0.7	77±0.1	75±1	69±0.4	64±0.3
15	VRB	68±0.3	76±0.6	77±0.1	81±0.3	65±0.1	73±0.1
16	VRG	70±0.8	96±0.4	75±0.8	76±0.9	61±1	60±0.9

4.5. Conclusion Chapter 4

Chapter 4 presented the attribute reduction method according to the Hausdorff topological approach. Experimental results show that the proposed algorithm is completely superior to other methods.

CONCLUDE

A. Main results of the thesis

Based on the set goals as presented in the introduction of the thesis, the main results of the thesis include: 1) Building an attribute reduction algorithm using the hybrid filter-wrapper approach using intuitive fuzzy distance (IFD) measurement. 2) Build an attribute reduction algorithm based on the filter approach (F_IFT) and the hybrid filter-wrapper algorithm (FW_IFT) using an intuitive fuzzy topology structure. 3) Build an attribute reduction algorithm according to the cluster filter-wrapper (CFW) hybrid approach using the Hausdorff topology structure. Experimental results on datasets downloaded from UCI show:

- The IFD algorithm can improve noise quite well, but the size and classification accuracy of the reduct set are less effective than the compared algorithms.

- The F_IFT algorithm has efficient execution time and a good reduced set size, but the classification accuracy is still limited compared to the compared algorithms.

- The FW_IFT algorithm for the reduced set has effective size and classification accuracy, but the execution time of the algorithm is limited compared to the compared algorithms.

- The CFW algorithm is entirely superior in execution time, size, and classification accuracy of the resulting reduct set is also superior to the best algorithms compared.

B. New contributions of the thesis

The research results of the thesis have contributed 03 attribute reduction methods, including:

- Attribute reduction method based on IFRS approach with new IF distance measure proposed.

- The reduction method belongs to the intuitive fuzzy topology approach based on new proposals about IF-subbase, IF-base, and unit topology.

- According to the Hausdorff topological approach, the attribute reduction method is based on new proposals about separability properties on the threshold fuzzy approximation space β.

C. Future development direction of the thesis

Incomplete decision tables with missing values appear quite commonly in data mining and machine learning. There have been many methods of reducing attributes in incomplete decision tables according to the extended rough set model approach. However, research results are still limited regarding the size and classification accuracy of the reduct. Therefore, the future research direction of the thesis will aim to reduce attributes for incomplete decision tables through several ways to expand the topological structure according to the rough set approach as follows: 1) Expand the topological structure based on the approximate space of the tolerance rough set model, study some separable properties to find the property selection criteria, and build the stopping condition of the algorithm. 2) Expand the topological structure based on the relationship of approximation operations of the tolerance rough set model, research some separable properties to find criteria for selecting properties, and build stopping conditions of the algorithm. 3) Develop incremental computing operations on topological space for dynamic data cases. 4) Develop the algebraic topology structure with new definitions of k-union operators and k-intersection operators to speed up the process of finding reduct sets.

LIST OF PUBLICATIONS

A. Published

[CT1] **Trần Thanh Đại**, Nguyễn Long Giang, Trần Thị Ngân, Hoàng Thị Minh Châu, "Rút gọn thuộc tính cho bảng quyết định đầy đủ theo tiếp cận Topo mờ", *Hội thảo quốc gia lần thứ XXIV: Một số vấn đề chọn lọc của Công nghệ thông tin và truyền thông*, Thái Nguyên, 12/2021 pp. 318-325, 2021.

[CT2] **Trần Thanh Đại**, Nguyễn Long Giang, Trần Thị Ngân, Hoàng Thị Minh Châu, Vũ Thu Uyên, Vương Trung Hiếu, "Về một phương pháp rút gọn thuộc tính cho bảng quyết định theo tiếp cận topo mờ trực cảm", *Các công trình nghiên cứu và phát triển CNTT và truyền thông*, Hà Nội, số 2, tr. 57-64, 2022.

[CT3] Nguyen Truong Thang, Nguyen Long Giang, **Tran Thanh Dai**, Nguyen Trung Tuan, Nguyen Quang Huy, Pham Viet Anh, Vu Duc Thi, "A Novel Filter-Wrapper Algorithm on Intuitionistic Fuzzy Set for Attribute Reduction from Decision Tables", *International Journal of Data Warehousing and Mining (IJDWM)* , số 17(4), tr. 67-100, 2021. (SCIE Q4 IF 0.78).

[CT4] **Trần Thanh Đại**, Nguyễn Long Giang, Hoàng Thị Minh Châu, Trần Thị Ngân, "Rút gọn thuộc tính cho bảng quyết định theo tiếp cận tập thô mờ trực cảm", *Kỷ yếu Hội nghị Khoa học Công nghệ Quốc Gia lần thứ XIII: Nghiên cứu cơ bản và ứng dụng công nghệ thông tin*, Nha Trang, 10/2020, tr. 516-524, 2020.

[CT5] **Trần Thanh Đại**, Nguyễn Long Giang, Vũ Đức Thi, Phan Đăng Hưng,"Về một phương pháp rút gọn thuộc tính theo tiếp cận tôpô Hausdorff", *Hội thảo quốc gia lần thứ XXVI: Một số vấn đề chọn lọc của Công nghệ thông tin và truyền thông*, Bắc Ninh, 10/2023,tr. 416-523, 2023.

B. Waiting review

[CT6] **Tran Thanh Dai**, Nguyen Long Giang, Vu Duc Thi, Tran Thi Ngan, Hoang Thi Minh Chau, Le Hoang Son "A New Approach for Attribute Reduction from Decision Table based on Intuitionistic Fuzzy Topology", *Soft Computing*. (SCIE Q2 IF 3.8). Đang chờ phản biện vòng 2.