

MINISTRY OF EDUCATION AND TRAINING
LAC HONG UNIVERSITY



DO SI TRUONG

**ATTRIBUTE REDUCTION METHOD AND DATA
CLUSTERING TECHNIQUE USING ROUGH SET
THEORY**

COMPUTER SCIENCE PHD THESIS ABSTRACT

Major: Computer Science

Code: 9480101

Dong Nai, 2023

The work is completed at: Lac Hong University

Science Instructor:

Associate Professor Nguyen Thanh Tung

Reviewer 1:

Reviewer 2:

Reviewer 3:.....

The thesis will be defended in front of the institutional examination council

.....

.....

At on.....

The thesis can be found at the library:

- The library of Lac Hong University
- National Library

TABLE OF CONTENTS

CHAPTER 1. INTRODUCTION	1
CHAPTER 2. OVERVIEW OF ROUGH SET THEORY AND ITS APPLICATION IN DATA MINING	3
2.1 Introduction	3
2.2 Some concepts of rough set theory	3
2.3 Related Concepts from Information Theory	5
2.4 Knowledge Discovery in Databases	6
2.5 Application of rough set theory in data mining	7
2.6 Conclusion	7
CHAPTER 3. FEATURE SELECTION USING ROUGH SET THEORY	8
3.1 Introduction	8
3.2 Feature selection	8
3.2.1 Subset generation	8
3.2.2 Subset Evaluation	9
3.3 Feature selection using rough set theory	9
3.4 Attribute Clustering Based Reduct Computing ACBRC	11
3.5 Conclusion	17
CHAPTER 4. CLUSTERING CATEGORICAL DATA	17
4.1 Introduction	17

4.2	Minimum Mean Normalized Variation of Information (MMNVI)	19
4.2.1	Basic Definitions and Idea of MMNVI	19
4.2.2	MMNVI Algorithm	20
4.2.3	Experimental Result	22
4.2.3.1	Bộ dữ liệu đánh giá	22
4.2.3.2	Performance Evaluation Methods	22
4.2.3.3	Clustering Results	22
4.3	Conclusion	23
CHAPTER 5. CONCLUSION AND FUTURE WORK		23
5.1	The results achieved and main contributions of this thesis	24
5.2	Future works	24

CHAPTER 1. INTRODUCTION

Knowledge Discovery in Databases (KDD) is a scientific field that researches and creates tools for discovering useful, hidden, and predictive knowledge within large databases [1, 2].

Research results and successful applications in recent times show that Knowledge Discovery in Databases is a potential scientific field, bringing many benefits and having numerous advantages compared to traditional data analysis tools. However, given the current rate of data growth, the research and application of data mining techniques are encountering numerous difficulties and challenges. This necessitates researchers to consistently strive in finding new tools to overcome these issues.

Rough set theory proposed by Z. Pawlak in the early 1980s [3], and is a relatively new soft computing tool for dealing with uncertain data. With innovative thinking, unique methodologies, and easy implementation, over the past three decades, the rough set theory has been researched, applied, and has become a crucial tool in the field of intelligent information processing [2, 4, 5, 6, 7, 8].

In that trend, many groups of scientists, including Vietnamese researchers, have been and are currently interested in attribute reduction in decision tables and data clustering. However, there are still some significant issues in this area of research that require further discussion and improvement. Therefore, I chose the research topic: “Attribute Reduction Method and Data Clustering Technique Using Rough Set Theory”.

The research objective of this thesis focuses on two problems within the topic: (1) propose new attribute reduction methods in a decision table; (2) discovering new algorithms for clustering categorical data using rough sets theory.

The research subject of the thesis is information systems and decision tables that may contain vague and uncertain data.

The scope of the thesis is researching data mining methods using rough set theory. Specifically, it focuses on two main issues outlined in the research objectives of the thesis.

Research methods: by synthesizing and evaluating research results using rough set theory in data mining published in both Vietnam and foreign scientific journals. On that basis, propose new techniques and algorithms, install, calculate, and compare and evaluate experimental results, proving the effectiveness of the algorithm. Following that, propose new techniques and algorithms, conduct installations, calculations, comparisons, and evaluations of test results, thereby demonstrating the effectiveness of the proposed algorithms.

The structure of this thesis includes an introductory chapter, three main content chapters, a concluding chapter, the author's published scientific articles, and a list of references. In Chapter 2, the basic concepts of rough set theory and some related concepts from information theory are presented. This chapter also provides an overview of data mining and explores the potential application of rough set theory in data mining. Chapter 3 presents the feature selection problem and some existing effective feature selection algorithms using the rough set theory, point out difficulties and challenges; on that basis, a reduct computing algorithm using attribute clustering is proposed. Chapter 4 discusses the data clustering problem in data mining, presents some existing effective data clustering techniques, points out their limitations, and proposes a new data clustering algorithm using rough set theory combined with concepts of entropy in information theory. Finally, the concluding chapter outlines the contributions of the thesis and suggests further developments.

The primary contributions of this thesis are outlined in Chapters 3 and 4.

Chapter 3 introduces a proposed algorithm for computing an approximate reduct in a decision table. It is called ACBRC (Attribute Clustering Based Reduct Computing). The experimental results show that the proposed algorithm is capable of computing approximate reduct with small size and high classification accuracy, when the number of clusters used to group the attributes is appropriately selected.

Chapter 4 proposes a new algorithm, called Minimum Mean Normalized Variation of Information (MMNVI). MMNVI algorithm uses the Mean Normalized Variation of Information of one attribute respect to another for finding best clustering attribute, and the entropy of equivalence classes

generated by selected clustering attribute for binary splitting the clustering dataset. Experimental results on real datasets from UCI indicate that MMNVI algorithm can be used successfully in clustering categorical data. It produces better or equivalent clustering results than the baseline algorithms.

The main contributions mentioned above were published in two articles in the Journal of Computer Science and Cybernetics in 2022 [CT1] and 2023 [CT2]. In addition, I'm also the author of an international article and co-author of three reports at prestigious domestic scientific conferences related to the thesis topic.

CHAPTER 2. OVERVIEW OF ROUGH SET THEORY AND ITS APPLICATION IN DATA MINING

2.1 Introduction

This chapter presents an overview of rough set theory, knowledge discovery in databases process and the applicability of rough set theory in data mining. The knowledge presented in this chapter is the foundation for researching new attribute reduction methods and new algorithms for clustering categorical data using rough sets theory.

2.2 Some concepts of rough set theory

Rough set theory proposed by Z. Pawlak [3], is a powerful mathematical tool for dealing with vague, imprecise, incomplete and uncertain data. This theory has been successfully applied in a number of areas such as machine learning, expert system, pattern recognition, and knowledge discovery in databases. This section presents some concepts of rough set theory.

Definition 2.1. [3] An information system is a pair $IS = (U, A)$, where U is a non-empty finite set of objects, A is a nonempty finite set of attributes, and for every $a \in A$ there is a mapping $a : U \rightarrow V_a$, where V_a denotes the domain of a .

Definition 2.2 [3]. Let $IS = (U, A, V, f)$ be an information system, $B \subseteq A$. Two elements $x, y \in U$ is said to be B -indiscernible in S if and only if $a(x) = a(y)$ for every $a \in B$.

We denote the indiscernibility relation induced by the set of attributes B by $IND(B)$. Obviously, $IND(B)$ is an equivalence relation and it induces unique partition (clustering) of U . The partition of U induced by $IND(B)$ in $IS = (U, A)$ denoted by $U/IND(B)$ or U/B and the equivalence class in the partition $U/IND(B)$ containing $x \in U$, denoted by $[x]_B$.

$$IND(B) = \{(u, v) \in U^2: a(u) = a(v) \ \forall a \in B\} \quad (2.1)$$

Definition 2.3 [3]. Let $IS = (U, A, V, f)$ and $X \subseteq U$. The B -lower approximation of X , denoted by $\underline{B}(X)$ and B -upper approximation of X , denoted by $\overline{B}(X)$, respectively, are defined by

$$\underline{B}X = \{u \in U: [u]_B \subseteq X\} \quad (2.2)$$

$$\overline{B}X = \{u \in U: [u]_B \cap X \neq \emptyset\} \quad (2.3)$$

Definition 2.4. [3] Let $IS = (U, A, V, f)$, $B \subseteq A$ and $X \subseteq U$. The accuracy of approximation of X with respect to B is defined as:

$$\alpha_B(X) = \frac{|\underline{B}X|}{|\overline{B}X|} \quad (2.4)$$

Throughout the thesis, $|X|$ denotes the cardinality of X .

Definition 2.5. [3] Let $IS = (U, A, V, f)$ be an information system, $B \subseteq A$ and $X \subseteq U$. The roughness of X with respect to B is defined as

$$R_B(X) = \alpha_B(X) = 1 - \frac{|\underline{B}(X)|}{|\overline{B}(X)|} \quad (2.5)$$

Definition 2.6. [3, 6] Decision table is an information system $DT = (U, C \cup \{d\})$, C is called the condition attribute set, while d is called the decision attribute.

Definition 2.7. [3, 6] Let $DT = (U, C \cup \{d\})$ be a decision table, $B \subseteq A$. The positive region of the d with respect to B , denoted by $POS_B(d)$, is defined as follows:

$$POS_B(d) = \bigcup_{X \in U/d} (\underline{B}X) \quad (2.6)$$

Definition 2.8. [3, 6] Given a decision table $DT = (U, C \cup \{d\})$, for any $B \subset C$, Pawlak defines the dependency degree of D on B in DT as follows.

$$\gamma_B(d) = \frac{|POS_B(d)|}{|U|}. \quad (2.7)$$

Obviously, there is $0 \leq \gamma_B(d) \leq 1$. If $\gamma_B(d) = 1$, then we say that D depends totally on B , and if $0 < \gamma_B(d) < 1$, then we say that d depends on B in a degree $\gamma_B(d)$. If $\gamma_B(d) = 0$, then we say that d does not depend on B .

2.3 Related Concepts from Information Theory

Given information system $IS = (U, A)$, and attribute $a \in A$. The IS information system can be viewed as a statistical population and attribute a as a discrete random variable. Suppose $V_a = \{x_1, x_2, \dots, x_h\}$, $U/\{a\} = \{X_1, X_2, \dots, X_h\}$. Then the probability distribution of a can be determined by

$$P(a = x_i) = P(x_i) = |X_i|/|U|, \quad i = 1, \dots, m. \quad (2.9)$$

Definition 2.9. [9] Let $IS = (U, A)$ be an information system, attribute $a \in A$. Shannon's entropy (entropy for short) of a is defined by the following expression:

$$H(a) = - \sum_{i=1}^m P(a = x_i) \log_2 P(a = x_i). \quad (2.10)$$

and by the convention $0 \log_2 0 = 0$.

Definition 2.10. [9] Let $S = (U, A)$ be an information system, $a, b \in A$. The join entropy $H(a, b)$ of a and b is defined by

$$H(a, b) = - \sum_{i=1}^m \sum_{j=1}^n P(a = x_i, b = y_j) \log_2 P(a = x_i, b = y_j). \quad (2.11)$$

The join entropy $H(a, b)$ is the measure of the amount of uncertainty two attributes a and b contain.

Definition 2.11. [9] Let $S = (U, A)$ be an information system, $a, b \in A$. The conditional entropy of a with respect to b denoted by $H(a|b)$ is defined as

$$H(a|b) = - \sum_{j=1}^n P(b = y_j) \sum_{i=1}^m P(a = x_i | b = y_j) \log_2 P(a = x_i | b = y_j). \quad (2.12)$$

The conditional entropy $H(a|b)$ quantifies the uncertainty of a random variable a when the outcome of another random variable b is known. We also have.

$$H(a|b) = H(a, b) - H(b) \quad (2.13)$$

Definition 2.12. [9] Let $S = (U, A)$ be an information system. The mutual information between the two attributes $a, b \in A$ is defined as

$$I(a; b) = H(a) - H(a|b) = H(b) - H(b|a) \quad (2.14)$$

Definition 2.13. [9, 10] Given an information system $IS = (U, A)$, $a_i, a_j \in A$. The normalized variation of information between a_i and a_j is defined by

$$NVI(a_i, a_j) = 1 - \frac{I(a_i; a_j)}{H(a_i, a_j)}. \quad (2.15)$$

$NVI(a_i, a_j)$ is a metric on the space of attributes, that is, for any attributes a_i, a_j , and a_k it satisfies

- (i) $NVI(a_i, a_j) \geq 0$ and the equality holds iff $a_i = a_j$,
- (ii) $NVI(a_i, a_j) = NVI(a_j, a_i)$,
- (iii) $NVI(a_i, a_j) + NVI(a_j, a_k) \geq NVI(a_i, a_k)$.

2.4 Knowledge Discovery in Databases

Knowledge Discovery in Databases (KDD) is a scientific field that researches and creates tools for discovering useful, hidden, and predictive knowledge within large databases [1, 2]. The following steps are included in KDD process: data selection, data preprocessing, data reduction, data mining,

knowledge representation. Among them, data mining is the most important problem. Data mining can be broadly categorized into two main types:

- Descriptive data mining focuses on summarizing and understanding the characteristics of historical data: Clustering, Summarization, Association rules.
- Predictive data mining involves analyzing current and historical data to forecast future events: Classification, Regression, Time-series analysis.

Among them, Data clustering (in the first group) and classification (in the second group) are two important techniques, widely utilized in the data mining process.

2.5 Application of rough set theory in data mining

Rough set theory can be applied to most stages of the process of KDD, including [5, 4, 6, 7, 11]

(1) Data preprocessing. In this process, rough set theory is used to reduce and clean the data.

(2) Data mining. In this process, rough set theory can be used to solve the following problems: data classification; data clustering; mining association rules.

Rough set theory is an effective tool for KDD. However, its applied research results to date still have many limitations.

2.6 Conclusion

This chapter includes 3 main parts: an overview of rough set theory with related concepts, the process of KDD and application of rough set theory in data mining. The basic concepts presented in this chapter are the basis for research and propose new attribute reduction methods in a decision table and new algorithms for clustering categorical data using rough sets theory.

CHAPTER 3. FEATURE SELECTION USING ROUGH SET THEORY

3.1 Introduction

This chapter provides an overview of feature selection, discusses the main methods to find an approximate reduct in a decision table, and introduces a new algorithm called ACBRC, based on attribute clustering method.

3.2 Feature selection

Feature selection can be viewed as the process of selecting a subset from the original set of features, removing as many irrelevant and redundant features as well as possible to improve the quality of data and reduce time and space complexity for analysis.

In general, a typical feature selection process consists of four basic steps [1, 12, 13]: subset generation, subset evaluation, stopping criterion, and result validation.

3.2.1 Subset generation

Subset generation is a process of heuristic search, with each state in the search space specifying a candidate subset for evaluation. The nature of this process is determined by two basic issues: decide the search starting point (or points) which in turn influences the search direction, and decide a search strategy.

Each newly generated subset needs to be evaluated by an evaluation criterion and compare it with the previous best subset. If the newly generated subset is better, it will replace the old one. A stopping criterion determines when the feature selection process should stop. Some frequently used stopping criteria are [12, 13]:

- The specified number of features has been collected;
- Some given bound is reached, where a bound can be a specified number (minimum number of features or maximum number of iterations);
- Subsequent addition (or deletion) of any feature does not produce a better subset;
- A sufficiently good subset is selected based on an evaluation criterion.

The final best subset must undergo verification by comparing data mining results using this "best" subset with the original features set on real or artificial datasets.

Another important issue in the feature selection process is determining how to evaluate the relevance of each subset.

3.2.2 Subset Evaluation

The goodness of a subset is always determined by a certain criterion. An evaluation criterion can be broadly categorized into two groups based on their dependency on mining algorithms that will finally be applied on the selected feature subset. We discuss the two groups of evaluation criteria below:

To evaluate a subset of features as optimal, it must be based on a certain criterion. A subset that is optimal according to one certain criterion is not necessarily optimal according to another certain criterion. An evaluation criteria can be broadly categorized into two groups based on their dependency on mining algorithms that will finally be applied on the selected feature subset: independent criteria and dependent criteria [12, 13].

Typically, an independent criterion is used in algorithms of the filter model. It tries to evaluate the goodness of a feature or feature subset by exploiting the intrinsic characteristics of the training data without involving any mining algorithm. Some popular independent criteria are distance measures, information measures, dependency measures, and consistency measures.

A dependent criterion used in the wrapper model requires a predetermined mining algorithm in feature selection and uses the performance of the mining algorithm applied to the selected subset to determine which features are selected. It usually delivers superior performance by identifying features better suited to the predetermined mining algorithm, but it also tends to be more computationally expensive and may not be suitable for other mining algorithms.

3.3 Feature selection using rough set theory

In the rough set theory, the process of feature subset selection in a decision table is viewed as reducts computation process. A reduct is a minimal subset of the conditional attributes set which provide the same information for

classification purposes as the entire set of available attributes. The classification task for the high dimensional dataset could be solved faster if a minimal reduct, instead of the original whole set of attributes, is used. Unfortunately, feature selection is considered as NP-hard problem, since the number of all subsets $2^n - 1$ grows exponentially with the number of features n . Searching for all reduced sets is only possible when n is relatively small. Therefore, several studies on computing approximate reducts have been carried out. An approximate reduct is a minimal reduct with acceptable errors, but can be found in much shorter time relative to an exact minimal reduct. In recent years, many approaches for computing approximate reducts were proposed [14, 15, 16, 1].

When dealing with high dimensional data (datasets with hundreds or thousands of features), many feature selection algorithms can successfully remove irrelevant features but fail to pull redundant ones out [14, 15, 17, 18, 16]. To overcome this problem, in the last decades some feature selection algorithms that use feature clustering were proposed in both supervised and unsupervised context [17, 18, 16, 19, 20, 21].

Clustering based feature selection algorithm follows the straightforward idea. It divides the initial feature space into a set of groups called clusters. Generally, correlation measures are used as clustering algorithm metrics, which makes features of the same group considered redundant. This leads to the selection of one feature to represent each cluster. The resulting feature subset is considered to be relevant and non redundant. However, for this type of approach, there are two core issues that need to be carefully considered, namely the choice of a similarity measurement function and a clustering method to use. The similarity function measures the similarity between two features in the feature space; clustering method collects features into groups using the selected similarity function.

In [17, 21], Hong et al. have built a similarity measure for a pair of attributes based on the relative dependency. Using this similarity measure, an algorithm called Most Neighbors First (MNF) is also proposed to cluster the attributes into a fixed number of groups. One of the big MNF deficiencies is that the convergence of the algorithm is not assured, in consequence it has to

be executed several times to find tendencies or patterns in the results.

Although clustering-based attribute reduction algorithms have received attention in recent times, their number is still relatively limited. In this chapter, the thesis proposes a clustering based attribute reduction algorithm for high dimensional decision table, named ACBRC.

3.4 Attribute Clustering Based Reduct Computing ACBRC

ACBRC algorithm works in three main stages. In the first stage, irrelevant attributes are eliminated. In the second stage relevant attributes are divided into a desired number of clusters by using Partitioning Around Medoids (PAM) clustering method integrated with a special metric in attribute space which is the normalized variation of information. In the third stage, the most representative attribute that is the most class-related is selected from each cluster to form a reduct. The proposed algorithm is implemented and experimented. The experimental results show that the proposed algorithm is capable of computing approximate reduct with small size and high classification accuracy, when the number of clusters used to group the attributes is appropriately selected.

In order to more precisely introduce the algorithm, we firstly present our definitions.

Definition 3.1. [1, 22] Let $DT = (U, C \cup \{d\})$ be a decision table. Attribute clustering in DT can be defined as the partitioning of the set C of conditional attributes into a collection $C_X = \{C_1, C_2, \dots, C_k\}$ of mutually disjoint subsets C_i of C , such that $C_1 \cup C_2 \cup \dots \cup C_k = C$, $C_i \neq \emptyset$, and $C_i \cap C_j = \emptyset$, for $i \neq j$.

Definition 3.2. Let $DT = (U, C \cup \{d\})$ be a decision table, $C \cup \{d\}$ is the full set of attributes. The distance between any pair of attributes X_i and X_j ($X_i, X_j \in C \cup \{d\}, i \neq j$) is measured by $NVI(X_i, X_j)$ defined as in (2.15)).

Note that for any $X_i \in C$ we have $0 \leq NVI(X_i, d) \leq 1$.

Definition 3.3. The irrelevance between the condition attribute $X_i \in C$ and the decision attribute d is measured by the distance value $NVI(X_i, d)$. The

greater the value $NVI(X_i, d)$ is, the lower the relevance between them. If $NVI(X_i, d)$ is greater than a threshold $\delta = 0.98$, we say that X_i is an irrelevant attribute; otherwise X_i is relevant one.

Definition 3.4. Let G be a cluster of attributes. A feature $X^R \in G$ is a representative attribute of the cluster if and only if

$$a^R = \operatorname{argmin}_{a \in G} NVI(a, d). \quad (3.1)$$

This means a^R is the strongest relevant attribute and can act as a relevant attribute for all the attributes in the cluster G .

Using the above definitions, ACBRC algorithm is the process consisting of the two connected parts: irrelevant attribute elimination and redundant attribute removal. The former obtains relevant attributes by eliminating irrelevant ones, and the latter removes redundant attributes from relevant ones via choosing representatives from different attribute clusters, and thus produces the final subset of attributes. Framework of ACBRC is shown in Figure 1.

Sử dụng các định nghĩa trên, thuật toán ACBRC là quá trình bao gồm ba công đoạn liên tiếp, với khung hoạt động được thể hiện trong hình 3.1.

ACBRC algorithm consists of three stages.

(1) First, irrelevant attributes are eliminated. For this purpose, the distance $NVI(a, d)$ is measured between each attribute d and the decision attribute d . We assume that the greater an attribute has irrelevance value is, the lower its ability to distinguish between classes. Here, the attribute with irrelevance greater than 0.98 will be removed from the initial attribute set.

2) Clustering the relevant attributes using function `pamk()` in R package “fpc”, integrated with metric NVI .

`pamk()` is an enhanced version of `pam()`, which can work with any distance matrix and does not require a user to choose the number of clusters k . Instead, it perform a PAM clustering with the number of clusters estimated by optimum average silhouette width method, described in [23].

(3) Finally, selecting from each cluster the attribute which has the strongest decision-relevance. This attribute can act as a representative attribute for all the attributes in the cluster. Once a attribute is selected, attributes

belonging to the same cluster are removed. The selected attributes form the approximate reduct.

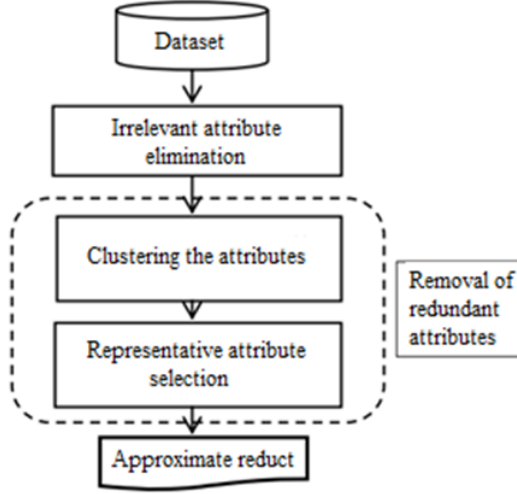


Figure 3.1 Framework of the proposed Reduct Computing algorithm ACBRC.

The main steps of the ACBRC algorithm are as follows:

Algorithm 3.1: ACBRC

inputs: The given decision table $DT = (U, C \cup \{d\})$, $\delta = 0.98$ - the irrelevance threshold.

output: Red - approximate attribute reduct.

step 1. Irrelevant attributes elimination.

For each $X \in C$ compute $irrelevance = NVI(X, d)$. If $irrelevance > \delta$ then $C^R = C \setminus \{X\}$.

step 2. Calculate the distance matrix NVI for all attribute pairs.

For each attribute pair X_i and X_j in C^R compute

$$NVI[i, j] = NVI(X_i, X_j) \quad (\text{equation (2.15)}).$$

step 3. Using `pamk()` fuction in R package "fpc" to cluster the attributes in C^R .

step 4. for each cluster G do $X^R = \operatorname{argmin}_{X \in G} NVI(X, d)$. $Red = Red \cup X^R$;

Experimental Results

The proposed ACBRC attribute reduction algorithm was implemented in R programming language and on a personal computer.

Experimental computations were carried out on 5 benchmark datasets obtained from UCI repository. The characteristics of these datasets are shown in Table 1. The first two columns show the names and abbreviations of datasets, the next two columns show the number of samples and attributes, and the last column shows the number of class labels. All attributes of the selected data sets are categorical. Thus, discretization is not necessary.

To evaluate the performance of our proposed ACBRC algorithm, we compare it with QuickReduct and CEBARKNC algorithms, in terms of the number of selected attributes, the execution time and the classification performance.

For comparing the classification performance of ACBRC, QuickReduct and CEBARKNC, we used the C5.0 and Native Bayes, which are two popular classification algorithms and widely applied in various research fields.

In order to make the best use of the data and obtain stable results, a 3-trials 10-fold cross-validation strategy is used. That is, for each data set, each attribute reduction algorithm and each classification algorithm, the 10-fold cross-validation is repeated 3 times, with each time the order of the instances of the dataset being randomized. Randomizing the order of the instances can help diminish the order effects. For each classification algorithm, we obtain 3-trials 10-fold classification accuracy for each attribute reduction algorithm and each dataset. Average these accuracies, we obtain mean accuracy of each classification algorithm under each attribute reduction algorithm and each data set. Experimental Result:

A. Comparison of number of selected attributes for three attribute reduction algorithms

Table 1 shows the attributes selected by three attribute reduction algorithms - ACBRC, QuickReduct and CEBARKNC when they are applied for each dataset.

TABLE 1. SELECTED ATTRIBUTES USING THREE ALGORITHMS

Datasets	ACBRC	QuickReduct	CEBARKNC
Chess	7 29 32 8 10 18 33 14 16 21	21 10 29 14 28 1 15 16 6 33 7 35 18 34 11 5 17 23 36 26 20 30 4 24 12 27 25 31 3 9 13	21 10 33 32 6 35 15 1 34 7 16 23 17 4 2 30 5 27 3 9 20 25 31 12 13 24 18 28 26 36
Mushroom	13 20 5 9	15 6 22 11 12 5 1	5 20 22 21
Soybean	22 21	22 1	22 4
Lung	9 48	1 42 7 4	9 43 3 4
Votes	4 12	1 9 14 4 11 16 13 3 2 6	4 11 3 13 16 2 9 15 1

From Table 1, we can see that all the three algorithms achieve significant reduction of dimensionality by selecting only a small portion of the original attributes. ACBRC generally obtains the best proportion of selected attributes.

TABLE 2. EXECUTION TIME OF THE PROPOSED ALGORITHMS (in sec.)

Datasets	ACBRC	QuickReduct	CEBARKNC
Chess	18.52	28.41	12.82
Mushroom	1.14	2.29	1.11
Soybean	0.64	0.12	0.36
Lung	0.84	0.31	0.44
Votes	0.64	0.73	0.53

From Table 2, we see that the execution time of the three algorithms depends on the characteristics of each dataset. In general, the execution time of the ACBRC algorithm is slightly larger than that of the QuickReduct and CEBARKNC algorithms, but the execution time of ACBRC is acceptable.

B. Evaluation of the classification performance of the ACBRC attribute reduction algorithm

Table 3 shows the 95% confidence intervals of 3-trials 10-fold classification accuracy of two classifiers on the 5 data sets without attribute reduction.

TABLE 3. CLASSIFICATION ACCURACY WITHOUT ATTRIBUTE REDUCTION

Datasets	C5.0	Bayes
Chess	0.9928 ± 0.0024	0.87868 ± 0.0126
Mushroom	1	0.94088 ± 0.0057
Soybean	0.975 ± 0.049	1

Lung	0.7667 ± 0.1701	0.56667 ± 0.2396
Votes	0.9674 ± 0.0139	0.90465 ± 0.0349

Table 4 shows the 95% confidence intervals of 3-trials 10-fold classification accuracy of two classifiers on the 5 data sets after ACBRC attribute reduction algorithm is used.

TABLE 4. CLASSIFICATION ACCURACY USING ATTRIBUTES SELECTED BY ACBRC

Datasets	C5.0	Bayes
Chess	0.9928 ± 0.0022	0.8906 ± 0.0129
Mushroom	1	0.9473 ± 0.0031
Soybean	1	1
Lung	0.8 ± 0.1996	0.6 ± 0.2134
Votes	0.9674 ± 0.0217	0.9581 ± 0.0164

Generally, for all five datasets, the classification accuracy of attributes selected by ACBRC is greater than the classification accuracy of original attributes.

C. Comparison of classification accuracy for three attribute reduction Algorithms

Table 5, and Table 6 show the 95% confidence intervals of classification accuracy by 3-trials 10-fold cross-validation when C5.0 and Naïve Bayes classifiers are used for the datasets with attributes selected by ACBRC, QuickReduct and CEBARKNC.

TABLE 5. C5.0 CLASSIFICATION ACCURACY USING DIFFERENT ATTRIBUTE REDUCTION ALGORITHMS

Datasets	ACBRC	QuickReduct	CEBARKNC
Chess	0.9928 ± 0.0022	0.9931 ± 0.0024	0.9937 ± 0.0035
Mushroom	1	1	1
Soybean	1	0.975 ± 0.049	0.975 ± 0.049
Lung	0.8 ± 0.1996	0.8 ± 0.1445	0.7667 ± 0.1701
Votes	0.9674 ± 0.0217	0.9651 ± 0.014	0.9558 ± 0.0126

From Table 5, we see that for datasets “Soybean”, “Lung”, and “Votes”, the C5.0 classification accuracy with attributes selected by ACBRC is greater than the classification accuracy with attributes selected by QuickReduct and CEBARKNC. And for

two other datasets, the classification result after attribute reduction by ACBRC is comparable with the results after attribute reduction by QuickReduct and CEBARKNC.

TABLE 6. BAYES CLASSIFICATION ACCURACY USING DIFFERENT FEATURE SELECTION METHODS

Datasets	ACBRC	QuickReduct	CEBARKNC
Chess	0.8906 ± 0.0072	0.8881 ± 0.015	0.8947 ± 0.0061
Mushroom	0.9473 ± 0.0036	0.982 ± 0.0027	0.9807 ± 0.0031
Soybean	1	0.875 ± 0.1096	1
Lung	0.5334 ± 0.2425	0.4667 ± 0.2789	0.4667 ± 0.2425
Votes	0.9581 ± 0.0164	0.9279 ± 0.023	0.9302 ± 0.0192

From Table 6, we see that for all five datasets, the Bayes classification accuracy with selected attributes by ACBRC is greater than the classification accuracy with attributes selected by QuickReduct and CEBARKNC.

3.5 Conclusion

These experimental results show that ACBRC is promising algorithm for attribute reduction.

CHAPTER 4. CLUSTERING CATEGORICAL DATA

4.1 Introduction

Clustering is a fundamental technique in data mining and machine learning. Let $D = \{x_1, x_2, \dots, x_n\}$ be the set of n objects, where each x_i is an N dimensional vector in the given feature space. The clustering activity is to find clusters/groups of objects in such a way that objects within the same cluster have a high degree of similarity, while objects belonging to different clusters have a high degree of dissimilarity [5]. Clustering problem appears in many different domains such as pattern recognition, information retrieval, computer vision, bioinformatics, medicine, etc. At present, there exist a large number of clustering algorithms in the literature [1].

Most of the earlier works on clustering have been focused on numerical data whose inherent geometric properties can be exploited to naturally define distance functions between data points. However, data mining applications frequently involve many datasets that consist of categorical attributes (such as gender, nationality, color, etc) on which there is no inherent distance measure

between categorical objects. Clustering categorical data is more challenging than clustering numerical data and clustering algorithms developed for numerical data cannot be used to cluster categorical data [1, 24, 25].

In recent years, clustering categorical data have attracted much attention from the data mining research community. Several algorithms for clustering categorical data have been proposed [1, 6, 7, 8, 26, 27]. These algorithms make important contributions to the problem of clustering categorical data, but they are not designed to handle uncertainty in the clustering process.

Rough set theory proposed by Z. Pawlak in the early 1980s [3], and is a relatively new soft computing tool for dealing with uncertain data. One of the major advantages of rough set theory is that it does not require any additional information about the data such as apriori probability distribution in statistics and membership function in fuzzy set theory [30]. In recent years, several rough set based divisive hierarchical clustering algorithms have been proposed for clustering categorical data [28, 25, 29, 30, 31]. The main idea of using rough set theory in clustering categorical data is to select a series of clustering attributes, where one of them is selected and used to split the objects at each time until all objects are clustered. Thus, the primary important task for this approach is to select from many candidates in a dataset one attribute that can best partition the objects.

The MMR algorithm was proposed by Parmar et al. In [32]. It is one of the most successful and pioneering rough set-based hierarchical clustering algorithms for categorical data. MMR algorithm determines the clustering attribute using roughness measure, which allows it to have the ability to deal with uncertainty.

Qin et al. [33] proposed an information-theory based divisive hierarchical clustering algorithm for categorical data that is implemented by selecting a clustering attribute using the mean gain ratio (MGR) and then selecting an equivalence class generated by the clustering attribute as one cluster using cluster entropy.

Although, Rough Set Theory based proposed categorical clustering algorithms make important contributions to the issue of clustering categorical data, they have some limitations such as they often have low accuracy and high

computational complexity. Especially, on some datasets they fail or hardly select their best clustering attribute [34].

In this chapter, we revisit two baseline divisive hierarchical clustering algorithms for use with categorical data and propose a new algorithm, called Minimum Mean Normalized Variation of Information (MMNVI). Two baseline algorithms are Min-Min Roughness (MMR), and Mean Gain Ratio (MGR). MMNVI iteratively performs only two steps on the current clustering dataset: (1) selecting a clustering attribute, (2) selecting an equivalence class generated by the selected clustering attribute as one cluster, and takes the union of other equivalence classes as the new clustering dataset. To implement the first step, MMNVI uses the concept of normalized variation of information in information theory which is a universal metric in the space of attributes. To perform the second step, MMNVI uses the concept of cluster entropy.

4.2 Minimum Mean Normalized Variation of Information (MMNVI)

4.2.1 Basic Definitions and Idea of MMNVI

As we have seen in Chapter 2, in a categorical information system $S = (U, A)$, each attribute in A defines a partition on the set U of objects. A good clustering of the objects should share as much information as possible with the partitions defined by each attribute in A [10, 35]. Also, the smaller the entropy of the data set, the more similar the objects in it are. Keeping this in mind, in our iterative clustering algorithm, at each iteration we expect to choose the clustering attribute whose partition is closest to those defined by other attributes; then in the partition defined by the chosen clustering attribute, select the equivalence class with the highest intra-class similarity as a cluster, and take the rest of the objects to form a new clustering dataset. The above two steps will be repeated on the new clustering dataset until the number of clusters obtained equals the pre-defined number k of clusters.

To measure the distance between two attributes, we use the NVI (Normalized Variation of Information) metric defined below.

Definition 4.1. Given an information system $IS = (U, A)$, $a_i \in A$. The Mean Normalized Variation of Information of a_i with respect to each $a_j \in A$, $a_j \neq a_i$, is defined by

$$MNVI(a_i) = \frac{1}{|A| - 1} \sum_{j=1, j \neq i}^{|A|} NVI(a_i, a_j). \quad (4.1)$$

Definition 4.2 Let $S = (U, A)$ be an information system, and $A = \{a_1, a_2, \dots, a_p\}$. Assuming the attributes in A are independent from each other, we define the entropy of dataset $X \subseteq U$ as following

$$Entropy(X) = H_X(a_1) + H_X(a_2) + \dots + H_X(a_p). \quad (4.2)$$

where $H_X(a_i)$ denotes the entropy of attribute a_i on X and is calculated by (2.6), $i = 1, \dots, p$. The smaller the entropy of X , the more similar the objects in X are. Therefore, the entropy of a cluster has been used by many authors as a measure to determine the intra-class similarity of a cluster [33, 36].

With the above idea and definitions, our MMNVI iterative algorithm works as follows:

4.2.2 MMNVI Algorithm

On the first iteration, it takes the set of all objects U as the clustering dataset and performs the following three main steps:

- Removes all the single-valued attributes.
- Select a clustering attribute as the one that has smallest MNVI value.
- Outputs as a cluster the equivalence class generated by the partitioning attribute which has lowest entropy, and takes as a new clustering dataset the union of other equivalence classes.

The above clustering process repeats until the number of the leaf nodes equals the pre-defined number k of clusters. Fig.1 shows the pseudo-code of the MMNVI algorithm.

Algorithm 4.1 MMNVI algorithm

Input: set of all objets U , set of all atributes A , the ginen number of clusters k .

Output: Clustering of U

Begin

Step 1: Set current number of cluster $CNC = 1$;

Set $CDataset = U$ // $CDataset$ denotes the actual clustering dataset

Step 2: $B = A$;

for each $a_i \in B$

determine partition $CDataset/Ind\{a_i\}$;

if $|CDataset/Ind\{a_i\}| = 1$ // all objects in $CDataset$ have the same values of a_i

$B = B - a_i$ //exclude attribute a_i ;

endif

endfor

Step 3: **for** each $a_i \in B$

calculate $MNVI(a_i)$ using formula (3.2) ;

endfor

Step 4: Determine the clustering attribute a^* which satisfies

$a^* = \operatorname{argmin}_{a_i \in B} MNVI(a_i)$;

Step 5: Determine partition $CDataset/Ind\{a^*\} = \{X_1, X_2, \dots, X_h\}$;

$X = \operatorname{argmin}_{X_i} (ent(X_i))$ for $X_i \in CDataset/Ind\{a^*\}$;

Step 6: output X as one cluster ;

$CNC = CNC + 1$;

if $CNC < k$

$CDataset = CDataset - X$;

go to Step 2 ;

else

output $CDataset$ as the last cluster;

endif

End

We have two following remarks about MMNVI algorithm:

(1) In Step 4, if there are many attributes having the same smallest MNVI value, we choose the first of them.

(2) In Steps 5 and 6, after having an attribute for splitting the clustering dataset, MMNVI selects the equivalence class with the lowest entropy as one cluster and takes the union of other equivalence classes as the new clustering

dataset. This is because the entropy of the second dataset is larger as indicated by the following proposition. In addition, if there are multiple equivalence classes with the same lowest entropy, we select the equivalence class with the largest number of objects.

4.2.3 Experimental Result

In order to test MMNVI, an implementation system was developed in R programming language, and tested on real-life data sets. Besides MMNVI algorithm, we also repeat two baseline algorithms, MMR and MGR, to compare with MMNVI.

4.2.3.1 Bộ dữ liệu đánh giá

To evaluate the clustering performance of MMNVI, we used 8 real-life data sets obtained from the UCI Machine Learning Repository, including Soybean small, Zoo, Votes, Breast Cancer Wisconsin, Mushroom, Balance Scale, Car Evaluation, and Chess.

4.2.3.2 Performance Evaluation Methods

To evaluate the performance of clustering algorithms, we use three widely used indexes. These indexes are: Overall Purity, Adjusted Rand Index (ARI), and Normalized Mutual Information (NMI).

4.2.3.3 Clustering Results

We first present the detailed clustering results of MMNVI, then we show the Overall Purity, Adjusted Rand Index and Normalized Mutual Information values obtained by MMR, MGR and MMNVI algorithms on 8 data sets.

Out of 8 data sets, MMNVI has the highest OP on the five data sets, specifically on the Soybean Small, Breast Cancer Wisconsin, Car evaluation, Votes and Balance scale. MMR has the highest OP on the Mushroom. MGR has the highest OP on the Soybean Small, Chess, and Zoo. On average, MMNVI achieves the highest Overall Purity.

With the ARI values of three algorithms, we see that MMNVI also has the highest ARI value on the four data sets, Breast Cancer Wisconsin, Votes, Chess and Balance scale. MMR has the lowest ARI value across all 8 data sets. MGR has the highest ARI value on Car evaluation, Mushroom, and Zoo. The

average ARI of each algorithm on 8 data sets. On average, MMNVI algorithm also achieves the highest ARI value.

Finally, the NMI values of three algorithms indicate the highest NMI value on the three data sets, Breast Cancer Wisconsin, Votes, and Balance scale. MMR has the highest NMI on the Soybean Small, Car evaluation, and Mushroom. MGR has the highest NMI on Chess, and Zoo. The average NMI of each algorithm on 8 data sets.

An important observation is that MMNVI do much better than other algorithms on the Breast Cancer, Votes, and Balance scale. Breast Cancer and Votes are data sets which have balanced class distribution. MGR do much better than other algorithms on the Zoo

In summary, MMNVI is a stable clustering algorithm and produces better or equivalent clustering results than the MMR and MGR algorithms.

4.3 Conclusion

In this paper, we have proposed a new divisive hierarchical clustering algorithm, called MMNVI (Minimum Mean Normalized Variation of Information), for categorical data. MMNVI algorithm uses the Mean Normalized Variation of Information of one attribute respect to another attributes for finding best clustering attribute, and the entropy of equivalence classes generated by selected clustering attribute for binary splitting the clustering dataset. MMNVI is easy to install. Experimental results on real-life data sets from UCI indicate that MMNVI algorithm can be used successfully in clustering categorical data.

CHAPTER 5. CONCLUSION AND FUTURE WORK

Knowledge Discovery in Databases (KDD) is a scientific field that researches and creates tools for discovering useful, hidden, and predictive knowledge within large databases. However, given the current rate of data growth, the research and application of data mining techniques are encountering numerous difficulties and challenges.

5.1 The results achieved and main contributions of this thesis

(1) By researching the algorithms proposed by researchers and finding shortcomings, the thesis has proposed a new algorithm for selecting attributes in a decision table based on clustering. ACBRC obtained the best proportion of selected attributes, and the best classification accuracy for C5.0, and Naive Bayes. The classification accuracy after attribute reduction by ACBRC even outperforms the classification accuracy using whole dataset in some cases. These experimental results show that ACBRC is promising algorithm for attribute reduction.

(2) By researching the algorithms proposed by researchers and finding shortcomings, the thesis has proposed a new algorithm, called Minimum Mean Normalized Variation of Information (MMNVI). Experimental results on real datasets from UCI indicate that MMNVI algorithm can be used successfully in clustering categorical data. It produces better or equivalent clustering results than the baseline algorithms.

The main contributions mentioned above were published in two articles in the Journal of Computer Science and Cybernetics in 2022 [CT2] and 2023 [CT3]. In addition, I'm also the author of an international article and co-author of three reports at prestigious domestic scientific conferences related to the thesis topic.

5.2 Future works

For the feature selection using rough set theory

(1) Research new attribute reduction algorithms in decision tables with missing values; (2) Applying the ACBRC algorithm to text classification

For clustering categorical data

(1) Enable the MMNVI algorithm to automatically discover the number of clusters. Instead of specifying the number of clusters, we can let the MMNVI algorithm stop splitting the clustering data set when the entropies of all the leaf nodes are lower than a predefined threshold value; (2) Extend MMNVI to handle both numerical and categorical data; (3) Perform more experiments on larger data sets with more objects and more attributes.

PUBLISHED WORKS

International Journal

[CT1] Do Si Truong, Nguyen Thanh Tung, Lam Thanh Hien, Improved minimum-minimum roughness algorithm for clustering categorical data, *International Journal of Advanced and Applied Sciences*, 8(10), Pages 43-50, 2021.

<https://doi.org/10.21833/ijaas.2021.10.006>

Vietnam Journal

[CT2] Do Si Truong, Lam Thanh Hien, Nguyen Thanh Tung, An effective algorithm for computing reducts decision table. *Journal of Computer Science and Cybernetics*, V.38, N.3 (2022), Pages 277–292, 2022.

DOI: <https://doi.org/10.15625/1813-9663/38/3/17450>.

[CT3] Do Si Truong, Lam Thanh Hien, Nguyen Thanh Tung, A new information theory based algorithm for clustering categorical data. *Journal of Computer Science and Cybernetics*, V.39, N.3 (2023), Pages 259-278, 2023.

DOI: <https://doi.org/10.15625/1813-9663/18568>

Vietnam Conference

[CT4] Pham Cong Xuyen, Do Si Truong, Nguyen Thanh Tung, An information-Theoretic metric based method for selecting clustering attribute. *Kỷ yếu Hội nghị Khoa học Quốc gia lần thứ IX “Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR’9)”*; Cần Thơ, ngày 4-5/8/2016, trang 31-40, 2016. DOI: 10.15625/vap.2016.0005.

[CT5] Nguyễn Minh Huy, Đỗ Sĩ Trường, Nguyễn Thanh Tùng, Về hàm đo độ phụ thuộc thuộc tính suy rộng. *Kỷ yếu Hội thảo quốc gia lần thứ XX: Một*

số vấn đề chọn lọc của Công nghệ thông tin và truyền thông, Quy Nhơn, 23-24/11/2017, trang 396-403, 2017.

[CT6] Đỗ Sĩ Trường, Trần Thanh Phương, Phạm Công Xuyên, Nguyễn Thanh Tùng, Thuật toán MMR cải tiến gom cụm dữ liệu phân loại, *Kỷ yếu Hội thảo “Fundamental and Applied IT Research 12/2021 (FAIR XIV)”*, Trường Đại học Công nghiệp Thực phẩm Tp.HCM, 2021.

REFERENCES

- [1] J. Han, J. Pei, H. Tong, "Data Mining: Concepts and Techniques," *Morgan Kanufmann*, vol. 3th Edition, 2012.
- [2] A. Skowron, S. Dutta, "Rough sets: past, present, and future," *Natural computing*, vol. 17(4), pp. 855-876, 2018.
- [3] Z. Pawlak, "Rough Sets - Theoretical Aspects of Reasoning about Data," *Kluwer Academic Publishers, Dordrecht*, 1991.
- [4] S. Pal, "Rough set and deep learning: Some concepts," *Academia Letters, 1849*, pp. 1-6, 2021.
- [5] P. Pieta, T. Szmuc, "Applications of rough sets in big data analysis: An overview," *International Journal of Applied Mathematics and Computer Science*, vol. 31, no. 4, p. 659–683 , 2021.
- [6] Q. Zhang, Q. Xie, G. Wang, "A survey on rough set theory and its applications," *CAAI Transactions on Intelligence Technology, 1*, pp. 323-333, 2016.
- [7] R. Bello, R. Falcon, "Rough sets in machine learning: a review," *Thriving Rough Sets*, pp. 87-118, 2017.
- [8] R. Słowiński, S. Greco, B. Matarazzo, "Rough-set-based decision support In Search Methodologies," *Springer, Boston, MA.*, pp. 557-609, 2014.
- [9] Y. Y. Yao, "Information-Theoretic Measures for Knowledge Discovery and Data Mining," *Part of the Studies in Fuzziness and Soft Computing book series (STUDFUZZ)*, vol. 119.
- [10] N. X. Vinh, J. Epps, J. Bailey, "Information Theoretic Measures for Clusterings Comparison: Variants, Properties Normalization and Correction for Chance," *Journal of Machine Learning Research*, vol. 11, pp. 2837-2854, 2012.
- [11] A. Skowron, C. Rauszer, "The Discernibility Matrices and Functions in Information Systems. Intelligent Decision Support,"

Handbook of Applications and Advances of the Rough Sets Theory, Kluwer, Dordrecht, p. 331–362, 1992.

- [12] U. Stańczyk, L. C. Jain, "Feature selection for data and pattern recognition: an introduction," *In Feature Selection for Data and Pattern Recognition*, vol. 584, pp. 1-7, 2015.
- [13] P. Dhal, C. Azad, "A comprehensive survey on feature selection in the various fields of machine learning," *Applied Intelligence*, pp. 1-39, 2021.
- [14] M. Alimoussa, A. Porebski, N. Vandenbroucke, R. O. H. Thami, S. El Fkihi, "Clustering-based Sequential Feature Selection Approach for High Dimensional Data Classification," *VISIGRAPP (4: VISAPP)*, pp. 122-132, 2021.
- [15] S. Chormunge , S. Jena, "Correlation based feature selection with clustering for high dimensional data.," *Journal of Electrical Systems and Information Technology*, 5, p. 542–549, 2018.
- [16] A. Janusz, D. Slezak, "Rough set methods for attribute clustering and selection," *Applied Artificial Intelligence*, vol. 28, p. 220–242, 2014.
- [17] T. P. Hong, Y. L. Liou, S. L. Wang, Bay Vo, "Feature selection and replacement by clustering attributes," *Vietnam J Comput Sci*, vol. 1, p. 47–55, 2014.
- [18] A. Janusz, D. Slezak, "Utilization of Attribute Clustering Methods for Scalable Computation of Reducts from High-Dimensional Data.," *Proceedings of the Federated Conference on Computer Science and Information Systems*, p. 295–302, 2012.
- [19] T. P. Hong, C. H. Chen, F. S. Lin, "Using group genetic algorithm to improve performance of attribute clustering, Appl.," *Soft Comput. J.*, p. <http://dx.doi.org/10.1016/j.asoc.2015.01.001>, 2015.
- [20] F. Pacheco, M. Cerrada, R. V. Sánchez, D. Cabrera, C. Li, J. V. de Oliveira, "Attribute clustering using rough set theory for feature selection in fault severity classification of rotating machinery.," *Expert Systems with Application*, pp. 69-86, 2017.

- [21] T. P. Hong, P. C. Wang, C.K. Ting, "An evolutionary attribute clustering and selection method based on feature similarity," *The IEEE Congress on Evolutionary Computation*, 2010.
- [22] G. Chandrashekar, F. Sahin, "A survey on feature selection methods," *Computers & Electrical Engineering*, vol. 40 (1), no. 40th-year commemorative issue, p. 16 – 28, 2014.
- [23] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53-65, 1987.
- [24] S. Naouali, S. B. Salem, Z. Chtourou, " Clustering Categorical Data: A Survey," *International Journal of Information Technology & Decision Making*, Vols. Vol. 19, No. 01, pp. 49-96, 2020.
- [25] J. Uddin, R. Ghazali, J. H. Abawajy, H. Shah, N. A. Husaini, A. Zeb, "Rough set based information theoretic approach for clustering uncertain categorical data," *PLOS ONE*, 2022.
- [26] G. Khandelwal, R. Sharma, "A simple yet fast clustering approach for categorical data.," *International Journal of Computer Applications*, 120(17)., pp. 25-30, 2015.
- [27] W. A. Hassanein, "Clustering algorithms for categorical data using concepts of significance and dependence of attributes," *European Scientific Journal*, vol. 10, pp. 381-400, 2014.
- [28] G. K. Singh, S. Mandal, "Cluster Analysis using Rough Set Theory," *Journal of Informatics and Mathematical Sciences*, Vols. 9, No. 3, pp. 509–520, , 2017.
- [29] S. Raheem, S. Al Shehabi, and A. M. Nassief, "MIGR: A Categorical Data Clustering Algorithm Based on Information Gain in Rough Set Theory," *International Journal of Uncertainty Fuzziness and Knowledge-Based Systems*, Vols. 30, No. 05, pp. 757-771, 2022.
- [30] T. Herawan, M. M. Deris, J. H. Abawajy, "A rough set approach for selecting clustering attribute," *Knowledge-Based Systems*, vol. 23, p. 220–231, 2010.

- [31] J. Uddin, R. Ghazali, M. M. Deris, "An empirical analysis of rough set categorical clustering techniques," *PloS one*, 12(1), 2017.
- [32] D. Parmar, T. Wu, J. Blackhurst, "MMR: an algorithm for clustering categorical data using rough set theory," *Data and Knowledge Engineering*, vol. 63, p. 879–893, 2007.
- [33] H. Qin, Xiuqin Ma, T. Herawan, J. M. Zain, "MGR: An information theory based hierarchical divisive clustering algorithm for categorical data," *Knowledge-Based Systems*, vol. 67, p. 401–411, 2014.
- [34] W. Wei, J. Liang, X. Guo, P. Song, Y. Sun, "Hierarchical division clustering framework for categorical data," *Neurocomputing*, vol. 341, p. 118–134, 2019.
- [35] Y. Y. Yao, Y. Zhao, J. Wang, "On reduct construction algorithms, Rough Sets and Knowledge Technology," *First International Conference, RSKT 2006, Proceedings, LNAI 4062*, pp. 297-304, 2006.
- [36] J. McCaffrey, "Data Clustering Using Entropy Minimization.," <http://visualstudiomagazine.com/Articles/2013/02/01/Data-Clustering-Using-Entropy-Minimization.aspx?Page=2&p=1>, 2018.
- [37] Mazlack, L. J., He, A., Zhu, Y., Coppock, S., "A rough set approach in choosing clustering attributes," *Proceedings of the ISCA 13th International Conference (CAINE 2000)*, p. 1–6, 2000.
- [38] Hoàng Thị Lan Giao, Khía cạnh đại số và logic phát hiện luật theo tiếp cận tập thô, Luận án Tiến sĩ Toán học, Viện Công Nghệ Thông Tin, 2007.
- [39] Nguyễn Đức Thuần, Phủ tập thô và độ đo đánh giá hiệu năng tập luật quyết định, Luận án Tiến sĩ Toán học, Viện Khoa học và Công nghệ Việt Nam, 2010.
- [40] Nguyễn Long Giang, Nghiên cứu một số phương pháp khai phá dữ liệu theo hướng tiếp cận tập thô, Luận án Tiến sĩ Toán học, Viện Công Nghệ Thông Tin, 2012.

- [41] J. Liang, K. S. Chin, C. Dang, R. C. M. Yam, "A new method for measuring uncertainty and fuzziness in rough set theory," *International Journal of General Systems*, vol. 31(4), pp. 331-342, 2002.
- [42] Y. Zhao, "R and Data Mining: Examples and Case Studies. <https://www.webpages.uidaho.edu/~stevell/517/RDataMining-book.pdf>," *Published by Elsevier*, December 2012.
- [43] M. J. Wierman, "Measuring uncertainty in rough set theory," *International Journal of General System*, 28(4-5), pp. 283-297, 1999.
- [44] J. Han , R. Sanchez , X. Hu, "Feature Selection Based on Relative Attribute Dependency: An Experimental Study," *Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing*, pp. 214-223, 2005.
- [45] J. Zhou, D. Miao, W. Pedrycz, H. Zhang, "Analysis of alternative objective functions for attribute reduction in complete decision tables," *Soft Comput* 15, p. 1601–1616, 2011.
- [46] Q. Song, J. Ni, G. Wang, "A fast clustering based feature subset selection algorithm for highdimensional data.," *IEEE Transactions on Knowledge and Data Engineering*, 25(1), p. 2013, 1–14.
- [47] K. Zhu, J. Yang, "A cluster-based sequential feature selection algorithm.," *In 2013 Ninth International Conference on Natural Computation (ICNC)*, p. 848–852, 2013.
- [48] C. Shannon , "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, pp. 379-423, 1948.
- [49] X. Hu, J. Han, T. Y. Lin, "A New Rough Sets Model Based on Database Systems," *Fundamenta Informaticae* 59 no. 2-3, pp. 135-152, 2004.
- [50] N. Xie, M. Liu, Z. Li, G. Zhang, "New measures of uncertainty for an interval-valued information system," *Information Sciences*, vol. 470, pp. 156-174, 2019.
- [51] M. Zhang, J. T. Yao, "A rough sets based approach to feature selection," *IEEE Annual Meeting of the Fuzzy Information*, 2004.

- [52] M. S. Raza, U. Qamar, "Feature selection using rough set-based direct dependency calculation by avoiding the positive region," *International Journal of Approximate Reasoning* 92, p. 175–197, 2018.
- [53] S. R. A. Ahmed, I. A. Barazanchi, Z. A. Jaaz, H. R. Abdulshaheed, "Clustering algorithms subjected to K-mean and gaussian mixture model on multidimensional data set," *Periodicals of Engineering and Natural Sciences (PEN)*, 7(2), pp. 448-457, 2019.
- [54] K. Thangavel, A. Pethalakshmi, "Dimensionality reduction based on rough set theory: A review," *Applied Soft Computing* 9, p. 1–12, 2009.
- [55] Z. Abbas, A. Burney, "A survey of software packages used for rough set analysis," *Journal of Computer and Communications*, 4 (9), pp. 10-18, 2016.
- [56] S. L. M. Belaidan, L. Y. Yee, N. A. Abd Rahman, K. S. Harun, "Implementing k-means clustering algorithm in collaborative trip advisory and planning system," *Periodicals of Engineering and Natural Sciences (PEN)*, 7(2), pp. 723-740, 2019.
- [57] B. K. Tripathy, A. Goyal, R. Chowdhury, A. S. Patra, "MMeMeR: An algorithm for clustering heterogeneous data using rough set theory," *International Journal of Intelligent Systems and Applications*, 9(8), 25., pp. 25-33, 2017.
- [58] R. Kohavi, G. H. John, "Wrappers for feature subset selection," *Artif. Intell.*, 97(1-2), pp. 273-324, 1997.
- [59] A. Jakulin, "Machine Learning Based on Attribute Interactions," *PhD Dissertation*, 2005.
- [60] A. Sandro, A. Pessoa, S. Stephany, "An Innovative Approach for Attribute Reduction in Rough Set Theory," *Intelligent Information Management*, pp. 223-239, 2014.
- [61] G. Y. Wang, H. Yu, D. C. Yang, "Decision table reduction based on conditional information entropy," *Chinese Journal Of Computers-Chinese* , vol. 25 (7), pp. 759-766, 2002.

- [62] D. Harris, A. V. Niekerk, "Feature clustering and ranking for selecting stable features from high dimensional remotely sensed data.," *International Journal of Remote Sensing*, 39(23), p. 8934–8949, 2018.
- [63] T. P. Hong, Y. L. Liou, "Attribute clustering in high dimensional feature spaces," *International Conference on Machine Learning and Cybernetics*, p. 19–22, 2007.
- [64] T. P. Hong, P. C. Wang, Y. C. Lee, "Cybernetics and Systems: An International Journal," *An effective attribute clustering approach for feature selection and replacement.*, p. 657–669, 2009.
- [65] D. E. Goldberg, "Genetic algorithms in search, optimization & machine learning," *Boston, MA, USA: Addison Wesley*, 1989.
- [66] L. Kaufman, P. J. Rousseeuw, "Computing groups in data: an introduction to cluster analysis," *John Wiley & Sons, Toronto*, 1990.
- [67] I. K. Park, G. S. Choi, "Rough set approach for clustering categorical data using information-theoretic dependency measure," *Information Systems*, vol. 48, pp. 289-295, 2015.
- [68] M. Halkidi, Y. Batistakis, M. Vazirgiannis, "On clustering validation techniques," *Journal of intelligent information systems*, vol. 17, pp. 107-145, 2001.
- [69] F. M. Reza, "An introduction to information theory," *Dover Publications, New York*, 1994.
- [70] H. Qin, Xiuqin Ma, J. M. Zain, T. Herawan, "A novel soft set approach in selecting clustering attribute," *Knowledge-Based Systems*, vol. 36, p. 139–149, 2012.
- [71] n.t. truong, "aaa," *abc*, 2023.
- [72] Pal, S., "Rough set and deep learning: Some concepts," *Academia Letters*, 1849, pp. 1-6, 2021.
- [73] Han, J., Pei, J., Tong, H, "Data Mining: Concepts and Techniques," *Morgan Kanufmann*, 2022.

- [74] X. Zhu, Y. Wang, Y. Li, Y. Tan, G. Wang, Q. Song, "A new unsupervised feature selection algorithm using similarity-based feature clustering," *Computational Intelligence*, vol. 35 (1), p. 2–22, 2019.
- [75] S. Mesakar, M. S. Chaudhari, "Review Paper On Data Clustering Of Categorical Data," *International Journal of Engineering Research & Technology*, vol. 1, no. 10, December, 2012.
- [76] J. Uddin, Rozaida Ghazali, Mustafa Mat Deris, "An empirical analysis of rough set categorical clustering techniques," *PLOS ONE*, Vols. 12, No. 1, 2017.
- [77] M. M. Baroud, S. Z. M. Hashim, J. U. Ahsan, A. Zainal, "Positive region: An enhancement of partitioning attribute based rough set for categorical data," *Periodicals of Engineering and Natural Sciences*, Vols. Vol. 8, No. 4, December 2020.
- [78] S. Raheem, S. Al Shehabi, and A. M. Nassief, "MIGR: A Categorical Data Clustering Algorithm Based on Information Gain in Rough Set Theory," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, Vols. Vol. 30, No. 05, pp. 757-771, 2022.
- [79] R. Jensen, Q. Shen, "A Rough Set-Aided System for Sorting WWW Bookmarks," *Lecture Notes in Computer Science*, vol. 2198, 2001.
- [80] R. Jensen, Q. Shen, "Computational Intelligence and Feature Selection: Rough and Fuzzy Approaches," *Wiley-IEEE Press eBook*, 2008.
- [81] R. Jensen, Q. Shen, "New approaches to fuzzy-rough feature selection," *IEEE Trans. Fuzzy Syst.* 17 (4), p. 824–838, 2009.
- [82] T. Herawan, "Rough Set Approach for Categorical Data Clustering," *A thesis submitted in fulfillment of requirements for the award of the Doctor of Philosophy*, 2010.
- [83] M. M. Baroud, S. Z. m. Hashim, A. Zainul, J. Ahnad, "An New Algorithm-based Rough Set for Selecting Clustering Attribute in

Categorical Data," *6th International Conference on Advanced Computing & Commucation Systems (ICACCS)*, pp. 1358-1364, 2020.

- [84] B. Tripathy, A. Ghosh, "SDR: An algorithm for clustering categorical data using rough set theory," *Recent Advances in Intelligent Computational Systems, IEEE*, pp. 867-872, 2011.
- [85] T. Herawan, "Rough clustering for cancer data sets," *International Journal of Modern Physics: Conference Series*, vol. 09, pp. 240-258, 2012.
- [86] N. T. T. Hien, H. V. Nam, "A k-Means-Like Algorithm for Clustering Categorical Data Using an Information Theoretic-Based Dissimilarity Measure," *Foundations of Information and Knowledge Systems: 9th International Symposium, FoIKS, Linz, Austr*, 2016.