REINITIATE LEVEL GREAT DELUGE ALGORITHM WITH COMPOSITE NEIGHBORHOOD STRUCTURES FOR ROUGH SET ATTRIBUTE REDUCTION

FAYYAD TALAL SALEM BANIHANI

PROJECT SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF MASTER OF COMPUTER SCIENCE

FACULTY OF INFORMATION SCIENCE AND TECHNOLOGY
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI

2018

ALGORITMA MEMULAKAN SEMULA PARAS BANJIR BESAR DENGAN STRUKTUR KEJIRANAN KOMPOSIT UNTUK SET KASAR PENGURANGAN ATRIBUT

FAYYAD TALAL SALEM BANIHANI

PROJEK YANG DIKEMUKAKAN UNTUK MEMENUHI SEBAHAGIAN DARIPADA SYARAT MEMPEROLEHI IJAZAH SARJANA SAINS KOMPUTER

FACULTI TEKNOLOGI DAN SAINS MAKLUMAT
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI

2018

DECLARATION

I hereby declare that the work in this thesis is my own except for quotations and summaries which have been duly acknowledged.

02 July 2018 FAYYAD TALAL BANIHANI

GP04688

ACKNOWLEDGEMENT

I wish to start my acknowledgement to express my earnest gratitude, my sincere heartfelt thanks and to almighty Allah swt for his blessing and to enable me completing this hard and complex work that has resulted in a successful outcome.

A special acknowledgement is due to University Kebangsaan Malaysia for the supporting and for the facilities. I would like to express my heartfelt thanks, sincere gratitude and appreciation to Assoc. Prof. Dr Salwani Abdullah my supervisor, for her invaluable guidance, support, time and encouragement through all the stages of this work. Furthermore, I would like to thank all the professors and lecturers who helped me through my work.

I would like to take this opportunity to extend my heartfelt thanks to many people, in different situations, in various countries, who so generously contributed to the work presented in this thesis.

Simultaneously, profound gratitude goes to my parents, my wonderful wife and the whole of my family for being there for me, believing in me, and helping in every period of this master thesis work. Without their amazing support, love and warm passion, this master thesis work might not be possible. They are the most important people in my world and I dedicate this thesis to them.

Lastly but not least, many thanks to all of my friends and to everyone who has offered me any piece of help appraise consultation and encouragement. I salute you.

ABSTRACT

Attribute reduction problem is one of the crucial matters that has been the concern in the studies investigating the complexity of real-life data. Attribute reduction is related to a process of finding minimum attribute reduction which is usually considered as an NP-Hard optimisation problem. Optimisation algorithms are found as the effective method to solve the NP-Hard optimisation problem by finding the minimum attribute out of a large set of attributes in information systems. This method is well-known for its usefulness in data mining and knowledge discovery. The numerous studies conducted employing meta-heuristic methods in solving the problems of attribute reduction has encouraged this study to propose an improved one single meta-heuristic approach as well. This proposed approach, Reinitiate Level Great Deluge algorithm with composite neighborhood structures for Rough Set Attribute Reduction problem (RLGD-RSAR), is generated from basic Great Deluge (GD) algorithm acknowledged for its simplicity in the parameter-setting. RLGD-RSAR proposes an intelligent mechanism to manage and reinitiate the value of the 'level' by sensing the lack of improvement for a certain number of repetitions plus the use of composite three different neighbourhood structures. The test on RLGD-RSAR has been carried out on 18 public domain datasets which are available in UCI machine learning repository. RLGD-RSAR is able to achieve competitive results in comparison with other available meta-heuristic approaches in the literature.

ABSTRAK

Pengurangan atribut merupakan satu perkara penting yang menjadi perhatian dalam kajian tentang penyiasatan data sebenar yang kompleks. Pengurangan atribut berkaitan dengan proses mencari atribut minimum yang biasanya dikategorikan sebagai masalah pengoptimuman NP-sukar. Algoritma pengoptimuman merupakan kaedah yang didapati berkesan untuk menyelesaikan masalah pengoptimuman NP-sukar dengan mencari atribut minimum daripada keseluruhan atribut dalam sesebuah sistem maklumat. Pelbagai kajian telah dijalankan menggunakan kaedah meta-heuristik dalam menyelesaikan masalah pengurangan attribut ini. Kajian-kajian tersebut menjadi faktor pendorong dalam kajian ini dalam mencari satu pendekatan meta-heuristik yang lebih efektif. Pendekatan yang dicadangkan iaitu *Reinitiate Level Great Deluge Algorithm with Composite Neighborhood Structures for Rough Set Attribute Reduction* (RLGD-RSAR) dengan tiga struktur kejiranan yang berlainan. Ujian RLGD-RSAR telah dijalankan ke atas 18 dataset domain awam yang terdapat dalam repositori pembelajaran mesin UCI. RLGD-RSAR Berjaya menghasilkan keputusan yang kompetitif berbanding dengan pendekatan meta-heuristik yang sedia ada.

TABLE OF CONTENT

		Page
DECLARATION		**iii**
ACKNOWLEDGEMENT		**iv**
ABSTRACT		**v**
ABSTRAK		**vi**
TABLE OF CONTENT		**vii**
LIST OF TABLES		**ix**
LIST OF FIGURES		**x**
CHAPTER I	**INTRODUCTION**	
1.1	BACKGROUND AND MOTIVATION	1
1.2	PROBLEM STATEMENT AND RESEARCH QUESTIONS	2
1.3	RESEARCH OBJECTIVES	4
1.4	SCOPE OF THIS STUDY	4
1.5	RESEARCH METHODOLOGY	4
1.6	OVERVIEW OF THIS STUDY	6
CHAPTER II	**A REVIEW OF ATTRIBUTE REDUCTION AND APPROACHES**	
2.1	INTRODUCTION	8
2.2	ATTRIBUTE REDUCTION (AR)	8
2.3	ROUGH SET THEORY: FUNDAMENTAL CONCEPTS	10
2.4	APROACHES ON ATTRIBUTE REDUCTION PROBLEMS	16
	2.4.1 Single-Based Approaches	17
	2.4.2 Population-Based Approaches	30
CHAPTER III	**REINITIATE LEVEL GREAT DELUGE**	

	ALGORITHM WITH COMPOSITE NEIGHBORHOOD STRUCTURES FOR ROUGH SET ATTRIBUTE REDUCTION	
3.1	INTRODUCTION	42
3.2	DATASETS DESCRIBTIONS	43
3.3	RLGD-RSAR PHASES	47
	3.3.1 Parameter Initialization Phase	48
	3.3.2 Initial Solution Construction Phase	49
	3.3.3 Improvement Phase	50
CHAPTER IV	**RESULTS AND DISCUSSION**	
4.1	INTRODUCTION	56
4.2	COMPARISON WITH THE STATE OF ART TECHNIQUES	56
4.3	SUMMARY	58
CHAPTER V	**CONCLUSION AND FUTURE WORK**	
5.1	RESEARCH SUMMARY	59
5.2	CONTRIBUTIONS	60
5.3	FUTURE WORKS	61

LIST OF TABLES

Table No		Page
Table 2.1.	Information system U\W a b c d	12
Table 2.2.	Information system U\W after reduct	15
Table 2.3.	GD litreture	29
Table 3.1.	Datasets specifications	43
Table 3.2.	Parameter Initialization	48
Table 3.2. RLGD-RSAR	Differences and similarities between GD-RSAR and	55
Table 4.1.	Comparison of results	57

LIST OF FIGURES

Figure No		Page
Figure 1.1	Research methodology phases	5
Figure 2.1	Data mining process	9
Figure 2.2	Attribute reduction rough set theory	11
Figure 2.3	Simulated annealing algorithm	18
Figure 2.4	Tabu Search Algorithm	21
Figure 2.5	Great Deluge algorithm for maximization	24
Figure 2.6	Ant Colony algorithm	31
Figure 2.7	Genetic algorithm	34
Figure 2.8	Whale Optimization Algorithm	40
Figure 3.1	RLGD_RSAR phases	47
Figure 3.2	Initial Solution Representation	49
Figure 3.3	Initial Solution with count number of selected attributes and dependency degree	49
Figure 3.4	Flip one point randomly (1Flip-Neig)	50
Figure 3.5	Flip two points randomly (2Flip-Neig)	51
Figure 3.6	Flip three points randomly (3Flip-Neig)	51
Figure 3.7	The RLGD_RSAR algorithm	52
Figure 3.8	The reinitiate level RLGD_RSAR acceptance solutions diagram	54

LIST OF ABBREVIATIONS

Great Deluge	GD
Genetic Algorithm	GA
Simulated Annealing	SA
Attribute Reduction	AR
Tabu Search	TS
Whale Optimization Algorithm	WOA
Ant Colony Optimization	ACO

CHAPTER I

INTRODUCTION

1.1 BACKGROUND AND MOTIVATION

Vast development of information technology and the massive challenges in information management have resulted in the familiarity of the term ‘big or complex data.’ It is, in fact, one of the major disruptors in the enterprises; and it exists in almost all industries and sectors such as banking, economics, medicare, training, manufacturing and even government. Big data needs a large storage space; otherwise, it will fail to deliver knowledge directly to the company. For this reason, this has become a critical issue for enterprises.

The data themselves, before being able to support operations in any organizations, need to be transformed first into useful knowledge and information. The knowledge is essential for analysts and managers to make decisions; because of this importance, data mining techniques, algorithms, and data mining software are proposed to assist solving the problems. Meanwhile, this study concentrates more on the attribute reduction which can be one of the preprocessing techniques for any data mining processes such as classification.

Attribute reduction plays a vital role in scientific development, named as an NP-hard improvement issue (Polkowski & Skowron 1998). The function works by getting the least attribute set derived out of the huge dataset of attributes by taking out unrelated and repeated parameters. This process assists to modify data superiority by managing any abnormality and nebulousness may exist. One of the most popular theories in this field is the Rough Set Theory which had been introduced initially by Pawlak (1982). This theory highlights the provision of the estimation of a confusing

approach set up by dual sophisticated techniques named as the lower and upper approximations.

Lately, there have been a number of scholars proposing meta-heuristic procedures to sort out the attribute reduction problems; and therefore, this study also proposes an improved great deluge algorithm which is called RLGD_RSAR. This algorithm suggests a sophisticated process to manage the phase value which will be done by modifying an original Great Deluge (GD) Algorithm found by (Dueck 1993). GD is one of the trajectory met heuristic algorithms which accepts the solutions while modifies the superiority of the current solution. However, it likewise consents the worst solutions depending on “Level”. RLGD_RSAR is initiated so as to investigate the optimal feature subsets based on a rough set theory which will be evaluated on some UCI datasets (https://archive.ics.uci.edu/ml/datasets.html n.d).

1.2 PROBLEM STATEMENT AND RESEARCH QUESTIONS

Attribute reduction carries problematic processes in dealing with statistics such as data mining, data extraction and data classification. Attribute reduction aims to perform reduction of attributes if database from information system where a risk of the information loss is minimal (Pawlak 1982). In particular, the common attributes are unnecessary to the decision attribute; consequently, the modest way to overcome this issue is by eliminating unrelated and back up properties from the input feature set or known as ‘reducing dimensionality’ and then picking up a part of beneficial properties.

According to Polkowski & Skowron (1998) it is proven that finding all possible minimizations attribute reductions from a chosen system is an NP-hard problem. Accordingly, due to the rapid growing magnitude and convolution datasets in the globe, attribute reduction problems have attracted many researchers to consider stochastic methods to the information systems in order to gather the information in large datasets into smaller ones. However, there has been no exact algorithm which is suitable to optimally solve this issue; instead, it will help in obtaining approximate or only near an optimum solution to solve the problem. Although it has a bit shortcoming, these methods can be looked into and thus improvement can be made on

one of them to reduce the effort and the time to get more optimized results rather than the former searches.

For the purpose to achieve studies objectives, an investigation will be performed to improve one single meta-heuristic approach, RLGD_RSAR, by using the dependency degree function that is calculated by indiscernibility relation as a rough set theory application to solve the problem. This procedure examines within attribute subsection range by employing the estimated precision rate as a standard of subsection appropriateness. Thus, by employing a multi initiate value of the boundary “Level” structure, it is to set up a range of reconnoitring shape at a time the algorithm discovers the quest range during the search process. The advantage of reinitiating level value structure is to handle equilibrium the deficiency by applying one certain magnitude of determining acceptance of the solutions structure in the search area.

However, based on the literature, most of the approaches used a single solution strategy or a population solutions strategy approaches to solve feature selection problem. The illustrations of a single solution approach are Simulated Annealing (Rosario & Thangadurai 2015) and Tabu search (Y. Wang et al. 2009). Whale Optimization Algorithm (Mafarja & Mirjalili 2018), Genetic algorithm (Kouser & Priyam 2018) and Ant colony algorithm (Chen et al. 2010) are classified as a population solutions approach.

Research Questions

Depending on the previous discussion in the problem statement, the aim of this study is to answer the following research questions:

1. How can the RLGD_RSAR algorithm help to find a better and appropriate feature subset from the original set?
2. How can the method avoid from easily get stuck in local optima using reinitiate level approach?

1.3 RESEARCH OBJECTIVES

Key purposes of our proposed work:

i. To develop a reinitiate level great deluge algorithm called (RLGD_RSAR) with composite neighbourhood structures in order to avoid from easily get stuck from local optima for rough set attribute reduction problem.
ii. To conduct a comparative analysis of the presentation of the recommended method with the latest technology.

1.4 SCOPE OF THIS STUDY

This study is mainly focused on the modified GD algorithm called RLGD_RSAR to solve rough set attribute reduction problems on 18 benchmark datasets found in UCI machine learning repository utilised to calculate the presentation of the recommended procedure.

1.5 RESEARCH METHODOLOGY

Principally, there were around five phases of our proposed research work and can be outlined as follows: The first phase is the Literature Review (LR). The second phase is the data pre-processing, third phase will concern the initial solution construction process, and fourth phase will comprise of the solution improvement or development process whereby the final or fifth phase will form the evaluation process respectively. This block diagram can be portrayed clearly in Figure 1.1 which illustrates the block diagram of the five phases carried out this research work.

Figure 1.1 Research methodology phases

Each phase contains its own descriptive summary and grounded on the following:

Phase 1: The Literature Review (LR)

LR presents the fundamental studies in regards to the problem of attribute reduction process in which has been methodically assessed; rewritten for the objective of acquiring the clear view of the problem background, structure the problem solving and ultimately conducting a comparative analysis of the existing studies methods.

Phase 2: Data Pre-processing

Data pre-processing phase addresses the collection of the needed authentic datasets generated from UCI machine learning repository, convert them to a more organized layout. This will be occurred by compensating some missing data erratically referenced on various datasets. The theory which is originated from the identification of a formal modality called as “Rough set theory”. This type of setup layout will be effectively utilized in the next phase.

Phase 3: Initial Solution Construction

This phase attempts to build and establish a preliminary remedy or solution by employing a constructive heuristic by taking into account that our proposed research work hence is defined as a random constructive heuristic and to be applied in order to present a random initial solution.

Phase 4: Solution Improvement

Phase 4 introduces the proposed and yet improved algorithm (RLGD_RSAR) in order to measure rough set attribute reduction problem. This phase employs a C++ programming language to be utilized with algorithm RLGD_RSAR which commences with a random initial solution. Alternatively, in order to produce an original solution, the neighborhood structure was designated.

Phase 5: Evaluation Process

The key functionality of the evaluation phase is to conduct a comparison between the presentation of tour recommended method with its counterpart throughout having the state-of-art approaches.

1.6 OVERVIEW OF THIS STUDY

This thesis is divided into five chapters that are organized in the following manner:

Chapter I outline the research background, its motivation with a description of the problem statement, followed by research objective, then the research scope with its approach.

Chapter II conducts a thorough valuation of the data mining conceptual attainment along with as well as classification task and its attribute reduction problem. The chapter will as well cover the concepts of Rough set theory concepts, the attributes reduction utilizing rough set theory. Moreover, this chapter will present its discussion of previously published works within the problematic area of interest. The

chapter will ultimately condense and analyses the results as well as data which were generated by these methods.

Chapter III describes the proposed methodology in which does have the methodology components, the methodology schematic diagram and the implemented pseudo code model. RLGD_RSAR approach to attribute reduction problem has been used for improvement by utilizing the structure of reinitiate level with composite neighborhood structures in order to produce elucidation and looking for the most optimum solution for all valid solutions.

Chapter IV, on the other hand, discusses comprehensively the results that were obtained from Chapter III methodology. The chapter then conducts a comparative analysis of the proposed research work and the other counterpart's methods. The mechanism of such comparative analysis has been executed in such a manner that was realized the attributes number of minimal reducts was attained by the proposed method by associated number by computing the degree of the dependency.

Finally, the conclusion in its principle addresses novelty that this thesis research works has achieved. This will contain the instrumental suggestions for future work throughout this study which will be presented in Chapter V.

CHAPTER II

A REVIEW OF ATTRIBUTE REDUCTION AND APPROACHES

2.1 INTRODUCTION

Chapter 2 presents a survey on the feature selection investigation zone. The survey offers a summarised description of attribute reduction concepts, rough set theory, data mining, and background of some related approaches that have been carried out to date for attribute reduction problem.

Also, chapter 2 encompasses three main Sections. Section 2.2 presents a brief introduction to attribute reduction. This part also contains a graphical representation of the data mining process. Section 2.3 discusses the rough set theory mechanism on how to work with an information system example to calculate the dependency degree. A brief summary of the previous approaches on attribute reduction problem will be discussed in section 2.4.

2.2 ATTRIBUTE REDUCTION (AR)

Attribute reduction (AR) is considered as an important factor in information illustration and data mining. It is categorised as minimization problem. AR reduces the data addressable and computational cost, and also provides users with a clearer picture and visual examination of the data of interest.

There is a tendency that the present info inclines to be more complex than conventional data; this is the main reason why AR process is used. The advantages of AR in data analysis, data mining, and data classification include the generation of smaller data in capacity and similar to methodical findings as the novel demonstration.

Before the reduction, the pre-processing arranged steps data is needed so as to obtain an efficient handling period while reducing and mitigating the trouble of dimensionality. AR is carried out on a huge data for increasing the superiority of acquaintance unearthing for specific requirements. Additionally, the minimization occurring in the datasets totally depends on the relationship between the characteristics to provide a more optimal demonstration of the findings, the removal of information loss is still being taken into consideration (Mafarja & Abdullah 2015).

Figure 2.1 (Lin et al. 2011) shows the whole process of data mining which consists of several steps. The first step is the understanding of the domain and related knowledge which is followed by the second step including selection of a dataset which is based on the needs and aims. The third step in this process is preprocessing which consists of several substeps namely: the removal of noise and outliers, the selection of related information, the handling of missing value, as well as the transformation of the data into suitable form for mining. The following step is applying the data mining technique. This process involves the search for patterns of interest in a specific representation form such as classification, regression or clustering. The two last steps include the evaluation of the pattern to identify the related patterns and the presentation of the knowledge.

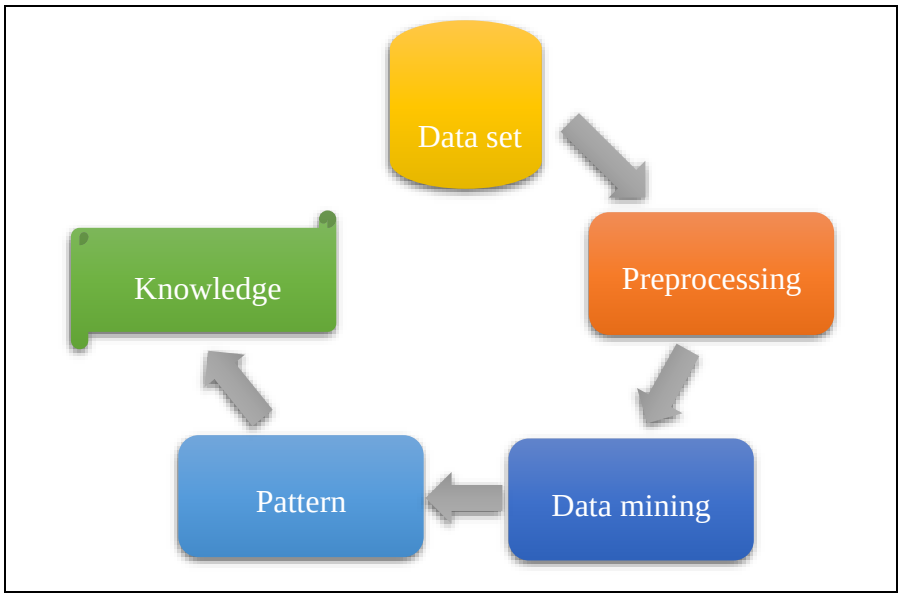

Figure 2.1 Data mining process

2.3 ROUGH SET THEORY: FUNDAMENTAL CONCEPTS

Basically, Rough Set Theory is obtained from key research on logical properties of information systems that are used as a scientific tool to treat the vague and the imprecise. From the outset, Rough Set Theory has been a methodology of database mining or information findings in interactive records. This section introduces the perceptions of Rough Set Theory; that concur relatively with other theories' assumptions that treat uncertain and vagueness information. Among the traditional approaches for the modelling and treatment of uncertainties that exist, there are two approaches i.e. theories of the uncertainty of Dempster-Shafer and Fuzzy Set that are closely related to Rough Set Theory (Pawlak et al. 1995).

Rough Set theory is defined as an effective mathematical method used to setup deficient data from information systems. Proposed by Pawlak (1982), Rough Set theory is the estimation of indistinct notion or set by a dual of accurate perception which is identified as a maximum and minimum approximations. This is the essential concept behind Rough Set theory. The subsection derived by minimum approximation is illustrated by objects that form part of an interesting subset; whereas the upper approximation is depicted by objects that will possibly form part of an interesting subset. Each subset which is defined through both upper and lower approximation is called as Rough Set. The Rough Set approach is also important for artificial intelligence and cognitive science especially in machine learning, knowledge discovery, data mining, expert systems, approximate reasoning and pattern recognition.

Because of the theory's merits, it has been used by numerous researchers and practitioners who essentially contribute to its progress and implementations. The Rough Set approach is easy, suits for concurrent processing and do not require somewhat further details with respect to the findings such as probability in statistics and membership rank in the fuzzy set theory. Besides, it offers well-organized procedures, algorithms and tools specially to find concealed shapes in findings; and it permits minimizing authentic findings – that is to find the lowest set of findings with the similar awareness available in its authentic findings. Other advantages including

the tolerance to evaluate the significance of data automatically generate the sets of decision rules from data and interpret obtained results straightforwardly. Figure 2.2 shows the graphical representation of AR procedure depends primarily on rough set theory (Tiwari et al. 2013).

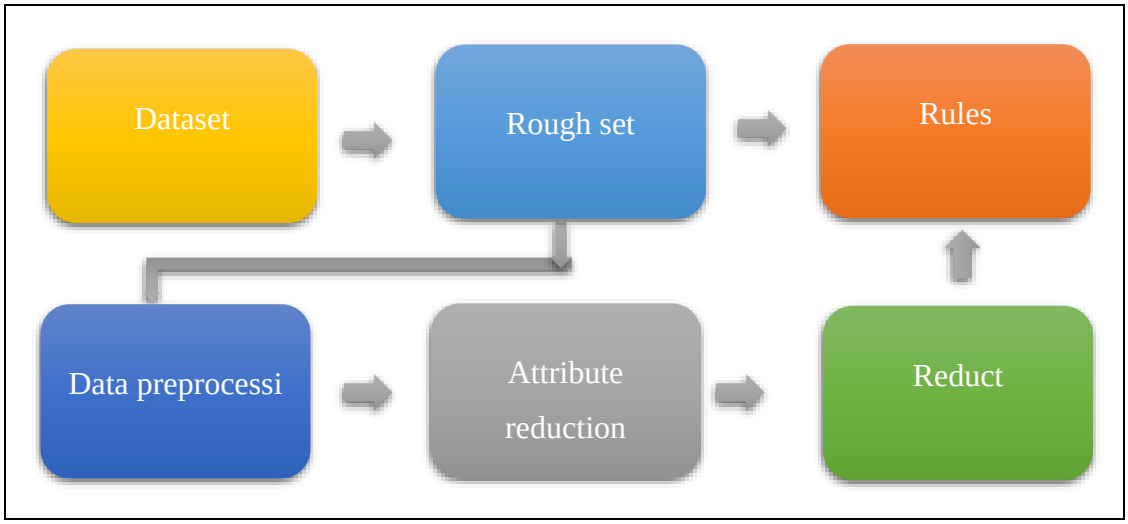

Figure 2.2 Attribute reduction rough set theory

Information System

Table 2.1 shows the said information system data encompasses of objects (rows) and attributes (columns). This table illustrated these data in which will be applied by the theory of Rough Set whereby every object has a specific amount of attributes (Lin 1997). The description of these objects are according to the format of the data table, in which rows represent objects for the analysis purposes while columns represent attributes (Wu et al. 2004). Let's assume that the Information System to be IS = (U, W), where U is a non-empty set of a finite object, A is a non-empty finite set of attributes that can monitor the rough set attribute reduction functionality.

Table 2.1 Information system U\W a b c d

U\W	*a*	*b*	*c*	*d*
S 1	0	1	1	0
S 2	1	0	0	0
S 3	1	1	1	1
S 4	0	1	1	0
S 5	1	0	1	1
S 6	0	0	0	0

Furthermore, Table 2.1 displays the information system representation example; hence it consists of two-dimensional array views example of a listed dataset, where the columns represent the attributes, while the rows represent the objects. Let's assume U represents the universe and W represents a set of attributes in the datasets. Let's consider C represents the set of condition attributes (a, b, c) and D represents the set of an attributes decision (d). Hence, C⊂ W, C∪ D=W, and C ∩ D=Ø. The entry in column x and row y has the value f (x, y), so f (x, y) defines equivalence relationship over U. Adding x, the universe can be partitioned into a set of disjoint subset:

$$R\,q = \{(x, y) \in U \wedge f(x, y) = f(x_0, y)\ \forall x_0 \in U\} \tag{2.1}$$

The intersection of all equivalence relations in *C* for any *C⊂W* is designated by *IND* (*C)* and is called as an indiscernibility relation over *C*. If *(x, y) ∈ IND(C)*, then *x* and *y* are indiscernible by attributes from *C*. The *IND*(*C*) relation can be written as:

$$IND\,(C) = \{(x, y) \in U^2 |\ \forall a \in C\ a\,(x) = a\,(y)\} \tag{2.2}$$

For example, let's consider *U= {S1, S2, S3, S4, S5, S6}, W= {a, b, c, d}, C= {a, b, c}* and *D= {d}*. The central idea for a rough set theory is the discernibility concept. Let's assume *IS= (U, W)* which is equal to be an information system, where *U* is a non-empty set of a finite object (the universe) and *W* is a non-empty finite set of attributes such that $a: U \rightarrow V_a$ for every $a \in W$. V_a represents the attribute value *a*. For any $P \subset W$ there is an associated equivalence relation *IND (P):*

If $(x, y) \in IND(P)$, then x and y are indiscernibility by the attributes in Table 2.1. While the equivalence classes of the P indiscernibility relation are denoted as $[x]p$.

For example, let $P = \{a, b\}$, then the indiscernibility relation can be partitioned into a set of disjoint subset:

$U/IND(P) = \{\{S1, S4\}, \{S2, S5\}, \{S3\}, \{S6\}\}$.

$U/IND(Q) = \{\{S1, S2, S4, S6\}, \{S3, S5\}\}$.

Let $W \subseteq U$. W can be calculated approximately using only the information contained within P by constructing the P-lower approximations of W:

$$\underline{P}W = \{x | [x]_p \subseteq W\} \tag{2.3}$$

And the P-upper approximation as follows:

$$\overline{P}W = \{x | [x]_p \cap W \neq \emptyset\} \tag{2.4}$$

By calculating the P-lower approximation as follows:

- If $W = \{S1, S2, S4, S6\}$ then $\underline{P}W = S1, S4, S6\}$

- If $W = \{S3, S5\}$ then $\underline{P}W = \{S3\}$

Let P and Q be an equivalence relation over U, then the positive regions can be defined as:

$$POSp(Q) = \bigcup_{W \in U/Q} \underline{P}W \tag{2.5}$$

The positive region contains all objects of U that can be classified as classes of U/Q using the information in attributes P. For example, let $P = (a, b)$ and $Q = (d)$, then:

$POSp(Q) = \cup \{\{S1, S4, S6\}, \{S3\}\} = \{S1, S3, S4, S6\}$

It is straightforwardly presented that the objects {*S1, S4, S6*} can only be categorized as belong to a class in attribute *d*, when considering attributes *a* and *b*, whilst the other objects cannot be classified as the information that make them discernible which is absent. Discovering dependencies between attributes is positioned as one of the key issues in RST. The dependency degree is computed according to the following formula:

$$k = \gamma_P(Q) = \frac{|POS\ P(Q)|}{|U|} \tag{2.6}$$

If $k = 1$, Q totally depends on P; if $0<k<1$, Q partially depends (in a degree of k) on P; and if $k = 0$ then Q does not depend on P. In the above example, the dependency degree of attribute $\{d\}$ from the attributes $\{a, b\}$ is calculated as:

$$k = \gamma_{\{a,b\}}(\{d\}) = \frac{|POS\{a,b\}(\{d\})|}{|U|} = \frac{|\{S1,S3,S4,S6\}|}{|\{S1\},\{S2\},\{S3\},\{S4\},\{S5\},\{S6\}|} = \frac{|4|}{|6|} = \frac{|2|}{|3|}$$

A reduct is defined as a subset of minimal Cardinality *Rmin* of the conditional attribute set C such $\gamma_{R(D)} = \gamma_{C(D)}$.

$$R = \{X: X \subseteq C, \gamma_{R(D)} = \gamma_{C(D)}\} \tag{2.7}$$

$$R_{min} = \{X: X \in R, \forall Y \in R, |X| \leq |Y|\} \tag{2.8}$$

The intersection of all the sets in Rmin called the Core:

$$Core(R) = \bigcap_{X \in R} X \tag{2.9}$$

The attributes elements cannot be eliminated without introducing more contradiction to the data set.

Using the dataset in Table 2.1, the dependency degree $D = \{d\}$ on all possible subsets of C can be calculated as:

$\gamma_{\{a\}}(\{d\}) = \frac{|3|}{|6|}$ $\gamma_{\{b\}}(\{d\}) = 0$ $\gamma_{\{c\}}(\{d\}) = \frac{|2|}{|6|}$

$\gamma_{\{a,b\}}(\{d\}) = \frac{|4|}{|6|}$ $\gamma_{\{a,c\}}(\{d\}) = 1$ $\gamma_{\{b,c\}}(\{d\}) = \frac{|3|}{|6|}$

$\gamma_{\{a,b,c\}}(\{d\}) = 1$

The minimal reduct set has identical dependency degree of the whole set for this particular example is:

$R_{\min} = \{a, c\}$ ……. $\gamma_{\{a,c\}}(\{d\}) = 1$

Table 2.2 Information system U\W after reducts

U\W	*a*	*c*	*d*
S 1	0	1	0
S 2	1	0	0
S 3	1	1	1
S 4	0	1	0
S 5	1	1	1
S 6	0	0	0

Based on the previous example, Table 2.2 demonstrates the information system after reduct one attribute (b) it can be established that the appropriate technique in order to get the most optimum solution is by scrutinizing all possible reducts available data. Wang et al. (2002) state that this method implementation on different problems will certainly setup more computational time due to the dataset size and considered as an NP-Hard problem. On the other hand, and according to Pawlak (1998) as cited by Kryszkiewicz & Lasek (2007) reduct would be possibly accommodating if the condition characteristics with lesser number are preferable rather than the classification accuracy.

According to the constraints of the time-consuming dispute, this method can only be applied to small size testing datasets. Consequently, a substitute strategy is required, and a well-organized heuristic with local search approach is deemed to have the capability of solving the problem.

The entry in column *x* and row *y* has the value *f (x, y)*, so *f (x, y)* defines an equivalence relation over *U*. Given *x*, the universe can be partitioned into a set of a disjoint subset (Jain et al. 2018).

The information system (IS) reduce is not measured singular. There is only one minimal reduct required for utmost requests even though there may be many subsets of attributes keeping the equivalence class structure described in the information system. To find such subset, it is indispensable to exhibit all potential subsets and make the selection by choosing the subsets with a maximum degree of dependency. This is a rather exorbitant procedure which can be considered because it can be applied only on small / simple dataset and there is only one reduct is required while the rest of the calculation won't be utilized. To obtain the solution, it is necessary to employ a shortcut method. In this method, the efficient algorithm with a solid structure can be used to generate the optimum solution by considering a maximum degree of dependency and minimum cardinality of reducts found. It means the present possible subset can be disregarded as it has less dependency degree or more elements.

2.4 APROACHES ON ATTRIBUTE REDUCTION PROBLEMS

Attribute reduction or feature selection is one of the processes in dealing with data like data mining, data extraction, and data classification. It has most of the problems and aims to perform reduction of attributes of a database from information systems where the information loss is minimal (Pawlak 1982). It has been proven that finding all possible minimal reductions from an information system is an NP-hard problem (Polkowski & Skowron 1998). Therefore, the AR problem attracted many researchers to consider stochastic methods to the information systems in order to accumulate the information in large datasets into smaller ones. Nevertheless, according to the previous research works, a majority of the procedure applied single solution procedure