A RELATIVE TOLERANCE RELATION OF ROUGH SET WITH REDUCT AND CORE APPROACH, AND APPLICATION TO INCOMPLETE INFORMATION SYSTEMS

RD. ROHMAT SAEDUDIN

A thesis is submitted in
fulfillment of the requirements for the award of the
Doctor of Philosophy

Faculty of Computer Science and Information Technology
Universiti Tun Hussein Onn Malaysia

OCTOBER 2020

ACKNOWLEDGEMENT

In the name of Allah, The Most Beneficent, The Most Merciful

All praises be to Allah, the Lord of the universe. May Allah bestow His mercy and grace upon His most beloved prophet Muhammad PBUH, his family, and his friends. My deepest gratitude to the grace of Allah SWT as with his bounty and mercy, then I can complete this Ph.D. thesis. I would like to take this opportunity to acknowledge the guidance and cooperation of my supervisors, examiners, colleagues, friends, and family members during this study period.

Primarily, I would express my sincere gratitude to my supervisor Prof. Dr. Mustafa Bin Mat Deris, for his continuous support, patience, encouragement and motivation, enthusiasm, and knowledge. Also the highest appreciation goes to my discussion partner, Iwan Tri Riyadi Yanto, Assoc. Professor Dr. Shahreen binti Kasim, Assoc. Professor Dr. Hairulnizam Mahdin, and Assoc. Professor Dr. Farhan md Fudzee for reminding me through his meticulousness, accuracy, dedication to work and insistence on perfection, and by his questions and comments. May Allah repay all of your kindness, Insyaa Allah.

I also would like to thanks to my parents Entin Martini and R Asikin, Encum Sumiati and Edi Sunardi, my lovely wife Siti Hajar Komariah, my sweet children Raden Daffa Muhammad Zahran, Raden Dafina Nur Azkiya Az-Zahra, and Raden Danish Muhammad Zahwan for their prayers, love, and encouragement. Thanks to everybody who contributed to this achievement directly or indirectly.

ABSTRACT

Data mining concepts and methods can be applied in various fields. Many methods have been proposed and one of those methods is the classical 'rough set theory' which is used to analyze the complete data. However, the Rough Set classical theory cannot overcome the incomplete data. The simplest method for operating an incomplete data is removing unknown objects. Besides, the continuation of Rough Set theory is called tolerance relation which is less convincing decision in terms of approximation. As a result, a similarity relation is proposed to improve the results obtained through a tolerance relation technique. However, when applying the similarity relation, little information will be lost. Therefore, a limited tolerance relation has been introduced. However, little information will also be lost as limited tolerance relation does not take into account the accuracy of the similarity between the two objects. Hence, this study proposed a new method called Relative Tolerance Relation of Rough Set with Reduct and Core (RTRS) which is based on limited tolerance relation that takes into account relative similarity precision between two objects. Several incomplete datasets have been used for data classification and comparison of our approach with existing baseline approaches, such as the Tolerance Relation, Limited Tolerance Relation, and Non-Symmetric Similarity Relations approaches are made based on two different scenarios. In the first scenario, the datasets are given the same weighting for all attributes. In the second scenario, each attribute is given a different weighting. Once the classification process is complete, the proposed approach will eliminate redundant attributes to develop an efficient reduce set and formulate the basic attribute specified in the incomplete information system. Several datasets have been tested and the rules generated from the proposes approach give better accuracy. Generally, the findings show that the RTRS method is better compared to the other methods as discussed in this study.

ABSTRAK

Konsep and kaedah perlombongan data boleh diaplikasikan dalam pelbagai bidang. Banyak kaedah telah dicadangkan dan satu daripadanya adalah '*rough set theory'* yang digunakan untuk menganalisis sistem maklumat lengkap. Namun, teori klasikal *Rough Set* tidak dapat mengatasi sistem maklumat tidak lengkap. Kaedah paling mudah untuk mengendalikan sistem maklumat tidak lengkap adalah untuk mengeluarkan objek-objek yang tidak diketahui. Selain itu, lanjutan kepada teori Rough Set yang dinamakan hubungan toleransi (*tolerance relation*) menghasilkan keputusan yang kurang menyakinkan dari segi penghampiran (*approximation*). Kemudian hubungan kesamaan (*similarity relation*) dicadangkan untuk memperbaiki keputusan yang diperolehi melalui teknik hubungan toleransi (*tolerance relation technique*). Walaubagaimanapun, apabila mengaplikasikan hubungan kesamaan (*similarity relation*) ini, sedikit maklumat akan hilang. Oleh itu, hubungan toleransi terhad (*limited tolerance relation*) telah diperkenalkan. Namun, sedikit maklumat juga turut akan hilang memandangkan hubungan toleransi terhad (*limited tolerance relation*) tidak mengambil kira ketepatan kesamaan antara dua objek. Justeru, kajian ini telah mencadangkan satu kaedah baru yang dinamakan Hubungan Tolerensi Relatif pada *Rough Set* yang berdasarkan hubungan tolerensi terhad (*limited tolerance relation*) yang mengambil kira kesamaan relatif ketepatan (*relative similarity precision*) di antara dua objek. Beberapa dataset tidak lengkap digunakan untuk pengkelasan dan perbandingan antara kaedah yang dicadangkan dengan kaedah-kaedah lain dilakukan berasaskan dua senario. Pendekatan ini memberi tumpuan kepada penghapusan atribut yang bertindan untuk menghasilkan set reduktif yang berkesan dan merumuskan set atribut utama bagi sistem maklumat yang tidak lengkap. Beberapa dataset telah diuji dan didapati bahawa, kaedah yang dihasilkan dari pendekatan ini dapat menghasilkan ketepatan yang lebih baik. Umumnya, penemuan menunjukkan bahawa kaedah RTRS lebih baik berbanding dengan kaedah-kaedah lain yang dibincang dalam kajian ini.

PUBLICATIONS

1. Saedudin, Rd Rohmat, et al. *Soft Set Approach for Clustering Graduated Dataset*. International Conference on Soft Computing and Data Mining. Springer, pages 631-637, Cham, 2016. (Indexed by Scopus)
2. Saedudin, Rd Rohmat, et al. *Rough Set Approach for Attribute Selection on Student Performance Dataset Based on Maximum Dependency Attribute*. The 5th International Conference on Electrical, Electronics, and Information Engineering (ICEEIE), pages 176-179, 2017. (Indexed by Scopus)
3. Saedudin, Rd Rohmat, et al. *A Comparative Analysis of Rough Sets for Incomplete Information System in Student Dataset*. “International Journal on Advanced Science, Engineering and Information Technology”, Vol 7 no 6 pages 2078-2084, December 2017. (Indexed by Scopus)
4. Saedudin, Rd Rohmat, et al. *A Relative Tolerance Relation of Rough Set for Incomplete Information Systems.* International Conference on Soft Computing and Data Mining. Springer, pages 72-81, Cham, 2018. (Indexed by Scopus)
5. Saedudin, Rd Rohmat, et al. *A Relative Limited Tolerance Relation of Rough Set for Potential Fish Yields in Indonesia*. Journal of Coastal Research, 84, pages 84-92, November 2018. (Indexed by Scopus and ISI)
6. Saedudin, Rd Rohmat, et al. *A Relative Tolerance Relation of Rough Set Approach and Its Application to Incomplete Information Systems*. Sains Malaysiana Journal,48(12), pages 2831-2839, December 2019. (Indexed by Scopus)

CONTENT

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

CPU	-	Computer personal unit
GHz	-	Gigahertz
GPA	-	Graphics Performance Accelerator
LTR	-	Limited Tolerance Relation
NLTRS	-	New Limited Tolerance Relation of Roush Set
NSSR	-	Non-Symmetric Similarity Relation
RTRS	-	Relative Tolerance Relation of Rough set
TR	-	Tolerance Relation
UCI	-	University of California, Irvine
US	-	United stated
VPRS	-	Variable Precission Rough Set

CHAPTER 1

INTRODUCTION

1.1 Background

Data mining is a field that deals with traditional data processing application software. Treats large, complex data sets to analyse matters, extract information systematically, or otherwise. Data with numerous cases (rows) offer more comprehensive statistical dynamism, while data with more extraordinary complexity (more characteristics or columns) may lead to a higher corrupt identification rate [1-3]. The challenges of data handling are capturing data, data storage, data summary, exploration, sharing, transfer, visualization, querying, updating, information privacy, and data source. Data is associated with three key concepts: volume, variety, and velocity.

Data mining is necessary for exploring knowledge or information derived from raw data, in which raw data is processed using software or selected methods to obtain information from the data. There are several processes involved in data mining, namely: selecting data, cleaning and pre-processing of data, reduction of data, data mining, and interpretation/evaluation.

In the fourth step in the knowledge discovery process or data mining is the process of extracting patterns from data. One of the methods of data mining is data clustering. Clustering in data mining is portioning a categorical data set into the homogenous class in a fundamental operation.

Pawlak was successful in proposing the rough set theory [4-6] in the study of intelligent systems characterized by uncertain or inconsistent data, especially in rule extraction [7-9], uncertainty reasoning [10-12], granular computing [13-16], data clustering [17-19], and data classification [20-22]. It has been proven an efficient mathematical tool compared to methods like principal component analysis (PCA), support vector machine, and neural networks [23-26]. In some methods, the Rough Set Theory counts the knowledge discovering process automatically based on the data without depending on prior Problem Statement knowledge [27-29]. The data used in all measurement by those methods were sample data.

Data sample is a set of objects collected or selected from a statistical population by a defined procedure. Sampling can be particularly useful with data sets that are too large to analyse in full -- for example, in data analytics applications or surveys. Identifying and analysing a representative sample is more efficient and cost-effective than surveying the entirety of the data or population. Many methods can be used to obtain or collect data. They are closed survey, open survey, interview, focus group discussion, and direct observation.

Surveys are particularly effective in collecting data, depending on structure and quality of survey questions. A survey is one of the common that data acquisition methods for data mining. In data, first mining can unusually find a study data set that contains total entries of each observation for all of the variables. Regularly, surveys and applications are often only somewhat completed by respondents. The reasonable analyses for incomplete data could be numerous, including indifference, deliberate avoidance of privacy, the ambiguity of the survey question, and aversion. Other causes that can make incomplete data are data integration, the data sampling technique used, and the fact of data is incomplete data.

In mining, a database with incomplete data, patterns of the missing data as the potential impacts of these missing data on the mining results constitute valuable information. There are two main approaches used to handle incomplete information systems. One is the second approach, which transforms the incomplete information system into a complete information system by replacing the missing attribute values with probable known characteristic values [30-33]. The other is the direct approach, which reaches the classical Rough Set Theory based on tolerance relations [34-37], similarity relations [38-41], and limited tolerance relations [42-45].

However, the tolerance relationship approach leads to poor results in terms of the approach. This bad result is caused by the absence of rules or conditions that require the similarity of the attribute entry values of the two objects being compared. Consequently, Stefanowski and Tsoukias [35, 36, 46, 47] introduced a similarity relationship to perfect the results obtained using a tolerance relationship approach. However, Wang et al., [37, 48-50] and Yang et al., [38, 51-53] showed that similarity relations would lose some information and so they proposed the relative tolerance relations. Nevertheless, some information might also be lost because of the limited tolerance relation does not consider the similarity precision between two objects. Nguyen et al., [39, 54-56] improved the tolerance relation by considering the probability of matching two objects. However, the probability distribution of data should be determined in advance.

However, the tolerance relation approach leads to poor results in terms of accuracy. In this research, the Relative Tolerance Relation based on Rough Set (RTRS) with core and reduct is proposed in order to get high accuracy with good response time.

1.2 Problem Statement

Nowadays, the increase in data is very huge in a digital era of more than 280 hexabytes. The huge data is only data without knowledge or information if it is not managed. Many researchers have analysed data to get knowledge and information, but they only took the complete data. Problems in the real world are statistics or empty attributes, incomplete statistics. The researcher knows that incomplete data would be difficult to process if the data still has some empty sections; hence, some data processing methods were used to fill the empty data groups. Some researchers proposed several methods to handle incomplete data i.e. tolerance relation, limited tolerance relation, and non-symmetric relation. To solve incomplete data in the existing method, the researchers used an imputation or removal of the incomplete data. The imputation methods used in those research are imputing missing data with mixed continuous and categorical variables, mean impute, and local least square. When the researcher imitates incomplete data, the data does not remain the same as the original data and the accuracy of the results becomes unsatisfactory. Moreover, by removing the incomplete data, the size of sample data will be smaller than before. What if the researcher has only limited data? He must collect data again to get more data, and if the researcher only analyses complete data, the accuracy of the result will be low too because of the small data sample.

In this study, the determination of the attribute weight value used a mathematical programming model for the incomplete data group. In this case, every missing value entry of an attribute will affect the weight value of an attribute. The distribution of weight values is resulted by the mathematical programming model.

Every attribute contributes to weight value in the data analysis. In the baseline method, data is processed by assuming that all attributes have the same contribution weight value in determining decisions. The problem arises in the real world when each attribute often has a different contribution weight value to the decision. It will be unfair if the attributes have the same weight value. The weight value process is important to understand each attribute weight value, so the researcher will know which are the lowest and the biggest weight value, so the lowest weight value attribute can be reduced.

The existing methods have handled the incomplete data in-group of data. However, the results of the existing method with data imputation or removing still show unsatisfactory result inaccuracy. The problem of accuracy will affect the knowledge of the result. The inaccurate knowledge or information can be a problem in making the decision.

1.3 Objectives

The main objective of this research is:

To propose a new method based on indiscernibility relation in handling incomplete data for better results of data classification.

This study has two sub objectives to achieve of the main objective, namely:

1. To propose an approach in handling incomplete data using improved limited tolerance relation based on indiscernibility relation for better accuracy.
2. To propose the improvement of weight value for improving processing time by reducing the attribute using reduct and core.

1.4 Scope and Limitation

The scope and limitation of this study falls within categorical data clustering using proposed methods based on the Rough Set Theory for incomplete data.

1.5 Contributions to the Study

The contributions of this study are:

1. A modified limited tolerance relation approach for categorical data clustering based on indiscernibility relation.
2. A related algorithm and proof of correctness of the proposed approaches
3. Comparative analysis and experiment results between clustering purity and the proposed approach with other baseline approach in terms of accuracy

1.6 Organization of this Study

This thesis is organized into seven chapters. A brief description of the contents for each chapter is given as follows:

1. Chapter 1 describes the challenges, problems, current methods, objectives, scopes, and significance of the study.
2. Chapter 2 describes the fundamental concept of the Rough Set Theory. The notion of an information system and its relation with a relational database, the concept of an indiscernibility relation induced by a subset of the whole set of attributes, the concept of an approximation space (Pawlak), the notion of set approximations and its quality of approximations are also described in this chapter. This chapter also explains the concept of a rough set in incomplete information systems i.e. tolerance relation, non-symmetric similarity relation, and limited tolerance relation.
3. Chapter 3 describes more about research methodology consists of conduct a literature review, collecting a real dataset related, propose a RTRS method to classify incomplete information systems, after that develop a program in MATLAB, apply the program to the data obtained, conduct results analysis, and results from validation.
4. Chapter 4 describes the proposed new approach of limited tolerance relation based on relative precision between two objects, namely relative tolerance relation of a rough set. This includes analysis, instrumentation, and data sources. Empirical studies based on seven benchmark datasets and real-world datasets demonstrate how the proposed method performs better compared to the rough set-based

methods. Furthermore, the application of the proposed method for clustering student data sets and marine data sets is presented. Discussion and analysis of the results of the proposed method are mentioned in detail here.

5. Chapter 5 describes enhanced RTRS using Variable Precision Rough Set (VPRS), reduct, and core to improve processing time in real data set.
6. Chapter 6 draws the general conclusions of the achieved results and presents the contributions together with a discussion of suggested topics for future studies.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This chapter strives to give a better understanding of the basic concepts of information systems and the rough set theory, before the extension of this theory in incomplete information systems i.e. tolerance relation, non-symmetric similarity relation, limited tolerance relation, and new limited tolerance relation [4-6,12]. The current research trends and directions are outlined before presenting the summary of the chapter.

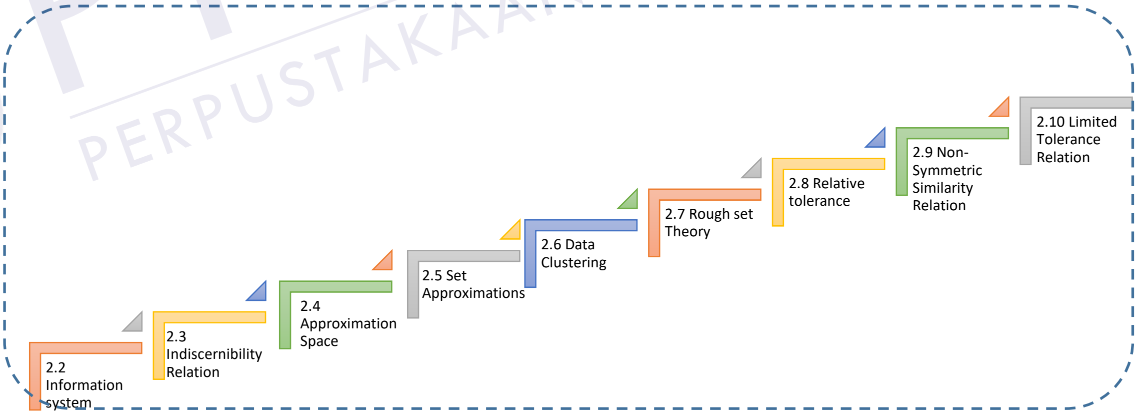

Figure 2.1: The content structure of chapter 2

2.2 Information System

The information system provides a convenient tool to represent the objects in terms of their attribute values. It is the 4-tuple (quadruple) $S = (U, A, V, f)$, where $U = \{u_1, u_2, \ldots, u_{|U|}\}$ is a non-empty finite set of objects, A = {a1 , a2 ,… , a|A| } is a non-empty finite set of attributes, $V = \bigcup_{a \in A} V_a$, *Va* is the domain (value set) of attribute a, f :U × A →V is an information function such that $f(u,a) \in V_a$, for every $(u, a) \in U \times A$, called information (knowledge) function [4, 12, 57-60]. An information system is also called a knowledge representation system or an attribute-valued system and can be intuitively expressed in terms of an information table (refer to Table 2.1).

Table 2.1: An information system

U	a_1	a_2	$\cdots$	A_k	$\cdots$	$A_{\|A\|}$
u_1	$f(u_1, a_1)$	$f(u_1, a_2)$	$\cdots$	$f(u_1, a_k)$	$\cdots$	$f(u_1, a_{\|A\|})$
u_2	$f(u_2, a_1)$	$f(u_2, a_2)$	$\cdots$	$f(u_2, a_k)$	$\cdots$	$f(u_2, a_{\|A\|})$
u_3	$f(u_3, a_1)$	$f(u_3, a_2)$	$\cdots$	$f(u_3, a_k)$	$\cdots$	$f(u_3, a_{\|A\|})$
.	.	.	$\ddots$	.	$\ddots$	.
u_U	$f(u_U, a_1)$	$f(u_U, a_2)$	$\cdots$	$f(u_U, a_k)$	$\cdots$	$f(u_U, a_{\|A\|})$

In many applications, there is a classification outcome. One (or more) distinguished attribute expressed this *a posteriori* knowledge called decision attribute. This process is known as supervised learning. A *decision system* is an information system of form *D* = (*U*, *A* = *C* U *D*, *V*, *f*) where *D* is the set of *decision attributes* and $C \cap D = \emptyset$. The elements of *C* are called *condition attributes*. A simple example of a decision system is given in Table 2.2.

A relational database is considered as an information system in which rows are labelled by the objects (entities), columns are labelled by attributes, and the entry in row *u* and column *a* has the value *f(u, a)*. It is noted that each map *f(u, a): U x A* → *V* is a tuple $t_i = (f\,(u_i, a_1),\ f\,(u_i, a_2),\ \ f\,(u_i, a_3),\ \ldots, f\,(u_i, a_{|A|}))$, for $1 \leq i \leq |U|$, where $|X|$ is the cardinality of X. Note that the tuplet is not necessarily associated with entity uniquely (refers to students 2 and 5 in Table 2.2). In an information table, two distinct entities could have the same tuple representation (duplicated/redundant tuple), which is *not permissible* in relational databases. Thus, the concepts in information systems are a generalization of the same concepts in relational databases.

Example 2.1 Data concerning 6 students, as shown in Table 2.2

Table 2.2: A student's decision system

Student	**Analysis**	**Algebra**	**Statistics**	**Decision**
1	Bad	Good	medium	accept
2	Good	Bad	medium	accept
3	Good	Good	Good	accept
4	Bad	Good	Bad	reject
5	Good	Bad	medium	reject
6	Bad	good	Good	accept

The following values were obtained from Table 2.2,

$U = \{1,2,3,4,5,6\}$,

$A = \{\text{Analysis, Algebra, Statistics, Decision}\}$, where

$C = \{\text{Analysis, Algebra, Statistics}\}$,

$D = \{\text{Decision}\}$

$V_{\text{Analysis}} = \{\text{bad, good}\}$.

2.3 Indiscernibility Relation

Table 2.2 showed that students 2, 3, and 5 are indiscernible (or similar or indistinguishable) concerning the attribute analysis. Meanwhile, students 3 and 6 are indiscernible concerning attributes algebra and statistics. Students 2 and 5 are indistinct for attributes analysis, algebra, and statistics. The starting point of the rough set theory

REFERENCES

[1] Breur, Tom (2016). Statistical Power Analysis and the contemporary "crisis" in social sciences. *Journal of Marketing Analytics*. 4 (2–3): 61–65]

[2] Hilbert, M. and López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. Science, 332 (6025), 60-65.

[3] Gilbert, D.T., King, G., Pettigrew, S. and Wilson T.D. (2016). *Comment on estimating the reproducibility of psychological science*. Science 351(6277): 1037.

[4] Z. Pawlak. (1982). Rough sets, *International Journal of Computer and Information Sciences*. vol. 11 (5), pp. 341-356,

[5] Orlowska, E. (Ed.) (1998). *Incomplete Information, Rough Sets Analysis. Physica-Verlag, Warsaw*. http://dx.doi.org/10.1007/978-3-7908-1888-8

[6] Skowron, A. and Suraj, Z. (2012). *Rough Sets and Intelligent Systems Professor ZdzisławPawlak in Memoriam*, Volumen 1. Springer Science & Business Media, Berlin

[7] Z. Lu and Z. Qin. (2011). Rule extraction from incomplete decision system based on novel dominance relation. *Proceedings of the 4th International conference on Intelligent Networks and Intelligent Systems*. pp. 149-152.

[8] Pandey SP, Krishnamachari A. (2006). *Computational analysis of plant RNA Pol-II promoters*. Biosystems. 83:38–50.

[9] Witten IH, Frank E. (2005). Data Mining: *Practical Machine Learning Tools and Techniques*. Morgan Kaufman; San Francisco. p. 560.

[10] J. Dai, W. Wang, Q.Xu, and H. Tian. (2012). Uncertainty measurement for interval-valued decision systems based on extended conditional entropy. *Knowledge-based Systems.* vol. 27, pp. 443-450.

[11] Guilong Liu. (2015). Special types of coverings and axiomatization of rough sets based on partial orders, *Knowledge-Based Systems*, v.85 n.C, p.316-321.

[12] Tutut Herawan, Mustafa Mat Deris, Jemal H. Abawajy. (2010). A rough set approach for selecting clustering attribute, *Knowledge-Based Systems,* v.23 n.3, p.220-231

[13] A. Skowron and P. Wasilewski. (2011). Toward interactive Rough-Granular Computing. *Control and Cybernetics*. vol. 40, no. 2, pp. 213-235.

[14] A. Skowron, J. Stepaniuk, and R. Swiniarski. (2010). Approximation spaces in Rough-Granular Computing. *Fundamentae Informaticae*. vol. 100, no. 1-4, pp. 141-157.

[15] Jankowski, J. and Skowron, A. (2009a). *Rough Granular Computing in Human-Centric Information Processing*. In: K.A. Cyran, S. Kozielski, J.F. Peters, U. Stańczyk and A. Wakulicz-Deja, eds., Man-machine Interactions. Springer, Heildelberg, 23-42.

[16] Jankowski, J. and Skowron, A. (2009b) Wisdom Technology: *A Rough-Granular Approach*. In: M. Marciniak and A. Mykowiecka, eds., Bolc Festschrift. LNCS 5070, Springer, Heildelberg, 3-41.

[17] I.T.R.Yanto, P. Vitasari, T.Herawan, and M.M.Deris. (2012). Applying variable precision rough set model for clustering suffering student's enxiety. *Expert Systems with Applications*. vol. 39(1), 452-459.

[18] T.Herawan, M.M.Deris, and J.H.Abawajy. (2009). A rough set approach for selecting clustering attributes. *Knowledge-Based Systems*. vol. 23(3), pp.220-231.

[19] D. Parmar, T. Wu, and J. Blackhurst. (2007). MMR: An algorithm for clustering categorical data using rough set theory. *Data & Knowledge Engineering*. vol. 63(3), pp. 879-893.

[20] D. Kim. (2001). Data classification based on tolerant rough set",*Pattern Recognition*. vol. 34(8), pp. 1613-1624.

[21] Kim, D., Bang, S.Y. (2000). *A handwritten numeral character classification using tolerant rough set*. IEEE transactions on PAMI 22, 923–937.

[22] Greco, S., Matarazzo, B., Slowinski, R. (2001). *Rough sets theory for multicriteria decision analysis*. European journal of operational research 129, 1–47.

[23] S. Trabelsi, Z. Elouedi, and P. Lingras. (2011). Classification systems based on rough sets under the belief function network. *International Journal of Approximate Reasoning*. vol.52 (9), pp.1409-1432.

[24] K. Kaneiwa. (2011). A rough set approach to multiple dataset analysis", *Journal of Applied Soft Computing*. vol. 11 (2), pp. 2538-2547.

[25] M. Bauer. (1997). *Approximations algorithm and decision making in the Dempster–Shafer theory of evidence – an empirical study*. IJAR, 17 (2–3) pp. 217-237

[26] Jiawei Han, Jian Pei, Yiwen Yin. (2000). *Mining frequent patterns without candidate generation, Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, p.1-12, May 15-18, Dallas, Texas, USA

[27] T. Yan and C. Han. (2014). A novel approach of rough conditional entropy-based attribute selection for incomplete decision system. *Mathematical Problems in Engineering*. vol. 2014, pp.1-15.

[28] S. Parsons, M. Kubat, and M. Dohnal. (1995). A rough set approach to reasoning under uncertainty. *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 7, no. 2, pp. 175–193, 1995.

[29] Z. Lu and Z. Qin. (2011). *Rule extraction from incomplete decision system based on novel dominance relation*. in Proceedings of the 4th International Conference on on Intelligent Networks and Intelligent Systems (ICINIS '11), pp. 149–152.

[30] M. Kryszkiewicz. (1998). Rough set approach to incomplete information systems. *Information Sciences*. vol. 112 (1-4),pp.39-49.

[31] Stefanowski, Jerzy. (1998). On rough set-based approaches to induction of decision rules". In Polkowski, Lech; Skowron, Andrzej (eds.). *Rough Sets in Knowledge Discovery 1*: Methodology and Applications. Heidelberg: Physica-Verlag. pp. 500–529.

[32] Yao, J. T.; Yao, Y. Y. (2002). Induction of classification rules by granular computing. *Proceedings of the Third International Conference on Rough*

Sets and Current Trends in Computing (TSCTC'02). London, UK: Springer-Verlag. pp. 331–338.

[33] Cornelis, C., De Cock, M. and Kerre, E. (2003). Intuitionistic fuzzy rough sets: *at the crossroads of imperfect knowledge, Expert Systems*, 20:5, pp260–270.

[34] M. Kryszkiewicz. (1999). Rules in incomplete information systems. *Information Sciences*. vol. 113 (3-4),pp.271-292.

[35] J. Stefanowski and A. Tsoukias. (1999). On the extension of rough set under incomplete information. *Lecture Notes in Artificial Intelligence*. vol. 1711.

[36] J. Stefanowski and A. Tsoukias. (2001). Incomplete information table and rough classification. *Computational Intelligence*. vol. 17 (3), pp. 545-566.

[37] G.Y. Wang. (2002). Extension of rough set under incomplete system. *IEEE International Conference on Fuzzy Systems.* pp. 1098-1103.

[38] X. Yang, X. Song, and X. Hu. (2011). Generalization of rough set for rule induction in incomplete system. *International Journal of Granular Computing, Rough sets and Intelligent Systems*. vol. 2 (1), pp. 37-50.

[39] D.V. Nguyen, K. Yamada, and M. Unehara. (2013). Extended tolerance relation to define a new rough set model in incomplete information systems. *Advances in Fuzzy Systems*. vol. 2013, pp. 1-11.

[40] Polkowski, et al. (2000). Rough Set Methods and Applications: *New Developments in Knowledge Discovery in Information Systems*, Studies in Fuzziness and Soft Computing. pp. 49-88

[41] W. Grzymala-Busse. (2004). Data with missing attribute values: *generalization of indiscernibility relation and rule induction*. Trans. Rough Sets I, pp. 78-95

[42] D. Parmar, T. Wu, and J. Blackhurst. (2007). MMR: an algorithm for clustering categorical data using rough set theory. *Data & Knowledge Engineering*. vol. 63, no. 3, pp. 879–893.

[43] Huang Z. (1998). Extensions to the k -Means Algorithm for Clustering Large Data Sets with Categorical Values. *Data Mining and Knowledge Discovery*. 2:283–304.

[44] Wu S, Liew AWc, Yan H, Member S, Yang M. (2004). *Cluster Analysis of Gene Expression Data Based on Self-Splitting and Merging Competitive*

Learning. IEEE Transactions on Information Technology in Biomedicine. 8(1):5–15.

[45] Hassanein W, Elmelegy A. (2013). An Algorithm for Selecting Clustering Attribute using Significance of Attributes. *International Journal of Database Theory & Application*. 6(5):53–66.

[46] Yawei Ge, Mingqing Xiao, Zhao Yang, Lei Zhang and Yajun Liang. (2018). *A hybrid hierarchical fault diagnosis method under the condition of incomplete decision information system, Applied Soft Computing*, (p350-365),

[47] Thinh Cao, Koichi Yamada, Muneyuki Unehara, Izumi Suzuki and Do Van Nguyen. (2017). Rough Set Model in Incomplete Decision Systems. *Journal of Advanced Computational Intelligence and Intelligent Informatics.* 21, 7, (1221).

[48] Y. Leung, W. Z. Wu, and W. X. Zhang, (2006). Knowledge acquisition in incomplete information systems*: A rough set approach.* European Journal of Operational Research, Vol. 168, pp. 164-180.

[49] W. Z. Wu, W. X. Zhang, and H. Z. Li, “Knowledge acquisition in incomplete fuzzy information systems via the rough set approach,” Expert Systems, Vol. 20, pp.280-286.

[50] M. Kryszkiewicz, (1999). Rules in incomplete information systems. *Information Sciences*, Vol. 113, pp. 271-292.

[51] HM Abu–Donia. (2012). Multi knowledge based rough approximations and applications. *Knowledge–Based Systems*, Vol. 26, pp. 20-29.

[52] HL Dou, XB Yang, JY Fan, and SP Xu. (2012). *The models of variable precision multigranulation rough sets*, in Rough Sets and Knowledge Technology–7th International Conference (Chengdu, China, 17–19 August, 2012), pp. 465-473.

[53] MA Khan and M Banerjee. (2008). *Formal reasoning with rough sets in multiple–source approximation systems*, International Journal of Approximate Reasoning, Vol. 49, pp. 466-477.

[54] J. Grzymala-Busse. (2004). Data with missing attribute values: *Generalization of indiscernibility relation and rule induction*. In J. Peters, A. Skowron, J. Grzymaa-Busse, B. Kostek, R. Winiarski, and M. Szczuka

(Eds.), Trans. on Rough Sets I, Vol.3100, Lecture Notes in Computer Science, pp. 78-95, Springer Berlin Heidelberg.

[55] J. W. Grzymala-Busse. (1991). On the unknown attribute values in learning from examples. In Z. Ras and M. Zemankova (Eds.), *Methodologies for Intelligent Systems*, Vol.542, Lecture Notes in Computer Science, pp. 368-377, Springer Berlin Heidelberg.

[56] M. Nakata and H. Sakai. (2007). *Handling missing values in terms of rough sets.* 23rd Fuzzy System Symp.

[57] J. Dai, W. Wang, Q. Xu, and H. Tian. (2012). Uncertainty measurement for interval-valued decision systems based on extended conditional entropy," *Knowledge-Based Systems*. vol. 27, pp. 443–450.

[58] Guilong Liu. (2015). Special types of coverings and axiomatization of rough sets based on partial orders, *Knowledge-Based Systems*, v.85 n.C, p.316-321.

[59] Guilong Liu , Zheng Hua , Zehua Chen. (2017). A general reduction algorithm for relation decision systems and its applications, *Knowledge-Based Systems*, v.119 n.C, p.87-93.

[60] S. Tsumoto. (2003). *Automated extraction of hierarchical decision rules from clinical databases using rough set model,* Expert Syst. Appl., p189-197.

[61] T. Herawan, M. M. Deris, and J. H. Abawajy. (2010). A rough set approach for selecting clustering attribute. *Knowledge-Based Systems*. vol. 23, no. 3, pp. 220–231.

[62] Molodtsov, D. (1999). Soft set theory-first results. *Computers and Mathematics with Applications 37*, 19–31.

[63] Roiger, R.J., Geatz, M.W. (2003). Data Mining: *A Tutorial-Based Primer. Addison Wesley,* Reading.

[64] Kong, Z., Gao, L., Wang, L., Li, S. (2008). *The normal parameter reduction of soft sets and its algorithm. Computers and Mathematics with Applications 56*, 3029–3037.

[65] J. Zhou and X. Yang. (2012). Rough set model based on hybrid tolerance relation. *International Conference on Rough sets and Knowledge Technology*. pp. 28–33.

[66] Qian, Y.H., Liang, J.Y., Li, D.Y., et al.. (2010). Approximation reduction in inconsistent incomplete decision tables. Knowledge-Based Systems 23, 427–433.

[67] Yang, X.B., Yang, J.Y. (2012). Incomplete information system and rough set theory: *models and attribute reductions.* Science Press Beijing & Springer

[68] E. Sutoyo, M. Mungad, S. Hamid, and T. Herawan, (2016). An efficient soft set-based approach for conflict analysis," PLoS One, vol. 11, no. 2.

[69] Q. Zhou. (2010). Research on Tolerance-Based Rough set Models. *System Science, Engineering Design and Manufacturing Informatization (ICSEM). 2010 International Conference on*. vol. 2, pp. 137–139.

[70] P. Mitra, C. A. Murthy, S. K. Pal. (2016). *Unsupervised Feature Selection Using Feature Similarity*. IEEE Trans. Pattern Anal. Mach. Intell. 24(3): 301-312.

[71] H. Yoshida, R. Leardi, K. Funatsu and K. Varmuza. (2012). *Feature selection by genetic algorithms for mass spectral classifiers, Anal. Chim.* Acta 446:485-494.

[72] Huang, Z. (1997). A fast clustering algorithm to cluster very large categorical data sets in data mining. *In Workshop on Research Issues on Data Mining and Knowledge Discovery*.

[73] Murty, M. N., Jain, A. K., & Flyn, P. J. (1999). Data clustering: A Review. *ACM Computing Surveys*. 31(2):264-323.

[74] Khaled S. Al-Sultana, M. Maroof Khan. (1996). *Computational experience on four algorithms for the hard clustering problem, Pattern Recognition Letter*s, v.17 n.3, p.295-308.

[75] Anderberg, M. R. (1973). *Cluster Analysis for Applications*. Academic Press, Inc., New York, NY.

[76] G. P. Babu, N. M. Murty, S. S. Keerthi. (2000). A stochastic connectionist approach for global optimization with application to pattern clustering, *IEEE Transactions on Systems, Man, and Cybernetics, Part B*: Cybernetics, v.30 n.1, p.10-24.

[77] Lelieveldt, B. P. F., Reiber, J. H. C., & Rezaee, R. (1998). A new cluster validity index for fuzzy C-Mean. *Pattern Recognition*. 19:237-246.

[78] Boudraa A.O. (1999). *Dynamic estimation of number of clusters in data sets. Electronics Letters 35*(19):1606–1607

[79] Theodoridis, K., & Koutroumbas, S., (1999), *Pattern Recognition, Academic Press.*

[80] Sripada, S. C., & Rao, S. M. (2011). Comparison of purity and entropy of KMeans clustering and fuzzy C-Means Clustering. *Indian Journal of Computer Science and Engineering*. 2(3):343-346.

[81] W. Husain, P. V. Low, L. K. Ng and Z. L. Ong. (2011). Application of Data Mining Techniques for Improving Software Engineering. *ICIT 2011 The 5th International Conference on Information Technology*.

[82] Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*. 27:379-423 and 623-659.

[83] Yeung, R. W. (2008). "*The Science of Information". Information Theory and Network Coding*. pp. 1–01.

[84] Zhao, Y., Karypis, G., & Fayyad, U. (2005). Hierarchical clustering algorithms for document datasets. *Data Mining and Knowledge Discovery*, 10(2):141-168.

[85] Beeferman, D. and Berger, A. (2000). Agglomerative clustering of a search engine query log. *In Proc. of the Sixth ACM SIGKDD Int'l Conference on Knowledge Discovery and Data Mining*, pp. 407–416.

[86] Boley, D. (1998). *Principal direction divisive partitioning. Data Mining and Knowledge Discovery*, 2(4):325–344.

[87] J. Stefanowski and A. Tsoukias. (1999). On the extension of rough set under incomplete information. *Lecture Notes in Artificial Intelligence*. vol. 1711.

[88] H. Midelfart, J. Komorowski, K. Nørsett, F. Yadetie, A. K. Sandvik, and A. Laegreid. (2002). *Learning rough set classifiers from gene expressions and clinical data.* Fundam. Inf., vol. 53, no. 2, pp. 155–183.

[89] Y. Sun, J. Han, P. Zhao, Z. Yin, H. Cheng, and T. Wu. (2009). Rankclus: integrating clustering with ranking for heterogeneous information network analysis," in EDBT '09: *Proceedings of the 12th International Conference on Extending Database Technology*. New York, NY, USA: ACM, 2009, pp. 565–576.

[90] T. Li, C. Ding, Y. Zhang, and B. Shao. (2008). Knowledge transformation from word space to document space,” in SIGIR ’08*: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval.* New York, NY, USA: ACM. pp. 187–194.

[91] P. Kolaitis and M. Vardi. (2007). A logical approach to constraint satisfaction. *In Finite Model Theory and Its Applications*, Springer 2007, pages 339–370.

[92] B. Rounds. (1991). Situation-theoretic aspects of databases. In Situation Theory and Appl., CSLI vol. 26. pages 229-256.

[93] P. Buneman, S. Davidson, A. Watters. (1991). A semantics for complex objects and approximate answers. JCSS 43, 170–218.

[94] J. Dy and C. Brodley. (2000). Future subset selection and order identification for unsupervised learning. Proc. 17th international conference machine learning.

[95] I. Kononenko. (1994). Estimationg attributes: *analysis and extension of relief.* Proc. Ninth International workshop machine learning pp. 171-182.

[96] M. Dash and H. Liu. (2000). Unsupervised future selection. *Proc. Pacific Asia Conference knowledge discovery and data mining.* pp. 110-121.

[97] Yuhua Qian, Jiye Liang, Witold Pedrycz, Chuangyin Dang. (2010). Positive approximation: *An accelerator for attribute reduction in rough set theory, Artificial Intelligence,* v.174 n.9-10, p.597-618.

[98] Fan Min, Huaping He, Yuhua Qian, William Zhu. (2011). *Test-cost-sensitive attribute reduction, Information Sciences:* an International Journal, v.181 n.22, p.4928-4942,

[99] Y. Nakahara, M. Sasaki, M. Gen. (1992). On the linear programming problems with interval coefficients, *Computers and Industrial Engineering, v.23* n.1-4, p.301-304.

[100] Xibei Yang, Dongjun Yu, Jingyu Yang, Lihua Wei. (2009). Dominance-based rough set approach to incomplete interval-valued information system, *Data & Knowledge Engineering,* v.68 n.11, p.1331-1347,

[101] Yee Leung, Manfred M. Fischer, Wei-Zhi Wu, Ju-Sheng Mi. (2008). A rough set approach for the discovery of classification rules in interval-

valued information systems, *International Journal of Approximate Reasoning,* v.47 n.2, p.233-246.

[102] M. Mat Deris, Z. Abdullah, R. Mamat, and Youwei Yuan (2015). A new limited tolerance relation for attribute selection in incomplete information systems", IEEE International Conference on Fuzzy Systems and Knowledge Discovery, 964-969.

[103] Saedudin, R. R., Sutoyo, E., Kasim, S., Mahdin, H., &amp; Yanto, I. T. R. (2017). A Comparative Analysis of Rough Sets for Incomplete Information System in Student Dataset. In International Journal on Advanced Science, Engineering and Information Technology (IJASEIT), 2017 Vol 7 No 6 (pp. 2078-2084).

[104] Anitha, K., Venkatesan, P. (2014). Rough Set Theory Approach to Generating Classification Rules. International Journal of Computational Intelligence and Informatics, Vol. 4 No. 3 (pp. 229-235)

[105] Vashist, Renu., Garg, M.L.(2011). Rule Generation Based on Reduct and Core: A Rough Set Approach. International Journal of Computer Application ,Vol. 29 No 9 (pp 1-5).