arity_Class_and_Enhanced_Tolerance_Relation_for_Complete_and.pdf

by

Submission date: 17-Apr-2023 06:08AM (UTC+0700)
Submission ID: 2066307744
File name: arity_Class_and_Enhanced_Tolerance_Relation_for_Complete_and.pdf (618.3K)
Word count: 4343
Character count: 21407

Data Reduction using Similarity Class and Enhanced Tolerance Relation for Complete and Incomplete Information Systems

Mustafa Mat Deris
Faculty of Computer Science and Information Technology
Universiti Tun Hussein Onn Malaysia
BatuPahat, 86400, Johor, Malaysia
mmustafa@uthm.edu.my

Mariam A. Hamid
Faculty of Computer Science and Information Technology
Universiti Tun Hussein Onn Malaysia
BatuPahat, 86400, Johor, Malaysia
mariamh@uthm.edu.my

Noraini Ibrahim
Faculty of Computer Science and Information Technology
Universiti Tun Hussein Onn Malaysia
BatuPahat, 86400, Johor, Malaysia
noraini@uthm.edu.my

Riswan Efendi
Department of Mathematics,
Universitas Islam Negeri Sultan Syarif Kasim
Riau, Indonesia
riswan.efendi@uin-suska.ac.id

Iwan Tri Riyadi Yanto
Universitas Ahmad Dahlan,
Yogyakarta, Indonesia
yanto.itr@is.uad.ac.id

Abstract— **Research on data analytics is entering a new challenge where huge data need to be processed in a timely manner. However, there are issues to computational resources when some data are redundant, inconsistent, noisy, and incomplete. It is imperative to reduce data size in order to overcome some issues specifically redundant and incomplete data. Thus, some redundant data or incomplete data need to be removed. In this paper, we would like to present the data reduction approaches using Rough set theory based on similarity class between the two objects and an enhanced tolerance relation when the data is incomplete/imprecise. The data structured with complete and incomplete information systems are discussed. Comparative analysis and experiment result between the proposed approaches to based-line approaches in terms of reduction rate and accuracy are presented. We found that, the proposed approaches are more favorable with high reduction rate for complete information systems and high accuracy for incomplete information systems.**

Keywords: *Data reduction, Similarity class, Enhanced tolerance relation*

I. INTRODUCTION

Information system has been very popular as it is commonly used for data representation to many entities. Business firms and other organizations rely on information systems in order to carry out and manage their business, such as managing human resources, processing financial accounts and to get new customers through online operations. To some fields, such as data mining [1], bio-informatics [2], and machine learning [3], information systems have huge data sets / data size and attributes that often be encountered. Some attributes/dimensions are irrelevant and some data are redundant that can complicate the problem and subsequently, degrade the performance and the accuracy. It is imperative to reduce data size and dimensionalities where it is a primary characteristic of massive data that is represented by the acquisition of storage spaces, data heterogeneity and also diverse dimensionalities. Data reduction is the process of reducing the amount of capacity required to store data Thus, some redundant data or irrelevant attributes need to be removed. Several popular methods/approaches for data reduction are PCA, neural networks (NN) and support vector machine (SVM) [4,5]. However, these approaches have some limitations: some variables are neglected when PCA is used; the size is still large when using NN, and SVM will optimize the parameters to model selection which is sensitive to over-fitting the model selection criterion [6] . Also, those approaches cannot cope with the incomplete information systems where some attribute values are missing.

Rough set theory [7] has been successfully used in analyzing information systems, especially in the study of rule extraction [8], uncertainty reasoning [9], granular computing [10,11], data clustering and data classification [12,13,14]. It is basically based on indiscernibility relation between objects. The approach is already proven and efficient as compared with PCA, neural networks and support vector machine [15,16] methods. Unlike those methods, rough set theory allows knowledge discovering process to be conducted automatically by the data themselves without any dependence on the prior knowledge [17].

In this paper, we would like to present the data reduction for structured and complete information systems using Rough set theory based on similarity class between the two objects and an enhanced tolerance relation for structured and incomplete information systems with accuracy. Comparative analysis and experiment result between the proposed approaches with tolerance relation approach in terms of accuracy are presented. We found that, the proposed approaches are more favorable and better in terms of accuracy.

978-1-7281-0045-6/19/$31.00 ©2019 IEEE

II. Theoretical Background

The concepts of information systems and similarity will be explained prior to the enhanced tolerance relation.

A. Information Systems

An information systems, S, is a 4-tuple (quadruple); $S=(U,A,V,f)$, where $U=\{u_1,u_2,\cdots,u_{|U|}\}$ denotes a non-empty finite set of objects and $A=\{a_1,a_2,\cdots,a_{|A|}\}$ denotes a finite set of attributes/ dimensions, $V=\bigcup_{a\in A}V_a$, where V_a is a value set of attribute a, $f:UxA\rightarrow V$ is a function such that $f(u,a)\in V_a$ for every $(u,a)\in UxA$ call information function [13]. If U in $S=(U,A,V,f)$ contains all objects with known values, the S is called *complete information system*, otherwise S has unknown or missing value and it is called *incomplete information system (IIS)*. The unknown value is denoted as "*" in incomplete information system. In this paper, we use the quadruple $S^{\#}=(U,A,V_{\#},f)$ to denote an incomplete information system. From the notion of an information system above, in the following sub-section we recall the notion of a similarity relation for complete information systems and tolerance relation as an approach for incomplete information system for data reduction.

B. Similarity Relations

Given a complete $S=(U,A,V,f)$, where $A=A^*\cup\{d\}$, A* is a set of condition attributes and d is a decision attribute, such that $f:UxA\rightarrow V$. For any subset $B\subseteq A^*$, the similarity relation, *Sim* is defined by the following definition [18,19].

Definition 1: *Let $S=(U,A,V,f)$, be a complete information system, and $B\subseteq A^*$. A similarity relation, Sim is defined as;*

$$\forall o_i,o_j\in U, Sim(o_i,o_j)\Leftrightarrow \forall a_k\in B,(a_k(o_i)=a_k(o_j))$$

C. Tolerance Relations

Given a complete decision system $S=(U,A,V,f)$, where $A=C\cup\{d\}$, C is a set of condition attributes and d the decision attribute, such that $f:U\times A\rightarrow V$, for any $a\in A$, where V_a is called domain of an attribute a. In incomplete information systems $S^B=(U,A,V_B,f)$, for any subset $B\subseteq C$, the tolerance relation T is defined by the following definition[20,21].

Definition 2. *Let $S^B=(U,A,V_B,f)$ be an IIS. A tolerance relation T is defined as*

$$\forall_{x,y\in U}\quad T(x,y)\Leftrightarrow \forall_{C_j\in B}\left(c_j(x)=c_j(y)\vee c_j(x)=*\vee c_j(y)=*\right).$$

Thus,

$$T=\{(x,y)x\in U\wedge y\in U\wedge \forall c_j(c_j\in B\rightarrow(c_j(x)=c_j(y)\vee c_j(x)=*\vee c_j(y)=*))\}$$

Obviously, T is reflexive and symmetric, but not transitive. From Definition 2, we describe the notion of tolerance class as follow.

Definition 3. *Let $S^B=(U,A,V_B,f)$ be an IIS. The tolerance class $I_B^T(x)$ of an object x with reference to an attribute set B is defined as $I_B^T(x)=\{y\mid y\in U\wedge T_B(x,y)\}$.*

From Definition 3, we describe the notion of lower and upper approximations of tolerance class as follow.

Definition 4. *Let $S^B=(U,A,V_B,f)$ be an IIS. The lower approximation x_B^T and upper approximation x_T^B of an object set X with reference to attribute set B respectively can be defined as follow* [22]*:*

$$x_B^T=\{x\mid x\in U\wedge I_B^T(x)\subseteq X\} \text{ and } x_T^B=\{x\mid x\in U\wedge I_B^T(x)\cap X\neq\phi\}$$

We can illustrate the above concepts with an IIS (for scholarship-application) below:

TABLE 1: AN INCOMPLETE INFORMATION TABLE

Students	C_1	C_2	C_3	C_4	Decision (d)
s_1	Good	Good	Fluent	*	*Accept*
s_2	Poor	*	Fluent	Good	*Accept*
s_3	*	*	Not fluent	Good	*Reject*
s_4	Good	*	Fluent	Good	*Accept*
s_5	*	Good	Fluent	Good	*Accept*
s_6	Poor	Good	Fluent	*	*Accept*
s_7	Poor	Good	Fluent	Good	*Accept*
s_8	*	Good	*	*	*Reject*
s_9	Good	Good	Fluent	Good	*Accept*

Example 1: Table 1, is a list of students $S=\{s_i|i=1,2,\ldots,9\}$ who apply for the scholarship sponsored by a Malaysian company. The decision is based on four criteria or condition attributes; the ability to do analysis (C_1), Studying BSc in Mathematics (C_2), the communication skills (C_3), and the ability to speak in Malay language (C_4). The table is an incomplete information system, where some values are not available, stated as '*'. The decision(d), where its domain values are *Accept* = $\{s_1,s_2,s_4,s_5,s_6,s_7,s_9\}$ and *Reject* = $\{s_3,s_8\}$. From Table 1, we will obtain the results by analyzing it with the tolerance relation in Definition 2, as follows:

$$I_C^T(s_1)=\{s_1,s_4,s_5,s_8,s_9\}, I_C^T(s_2)=\{s_2,s_5,s_6,s_7,s_8\},$$
$$I_C^T(s_3)=\{s_3,s_8\}, I_C^T(s_4)=\{s_1,s_4,s_5,s_8,s_9\},$$
$$I_C^T(s_5)=\{s_1,s_4,s_5,s_8,s_9\}, I_C^T(s_6)=\{s_2,s_5,s_6,s_7,s_8\},$$
$$I_C^T(s_7)=\{s_2,s_5,s_6,s_7,s_8\},$$

$I_C^T(s_8)=\{s_1,s_2,s_3,s_4,s_5,s_6,s_7,s_8,s_9\}$,
$I_C^T(s_9)=\{s_1,s_3,s_4,s_5,s_8,s_9\}$, and

$$\frac{U}{IND(d)}=\{\{s_1,s_2,s_4,s_5,s_6,s_7,s_9\},\{s_3,s_8\}\}$$

Thus,

$Accept_C^T=\phi$, $Reject_C^T=\{s_3,s_8\}$,
$Accept_T^C=\{s_1,s_2,s_4,s_5,s_6,s_7,s_8,s_9\}$,
$Reject_T^C=U$.

From the above analysis, some objects that can be discerned intuitively cannot be classified, such as s_7 and s_9 have complete information, but s_7 and s_9 are not in the lower approximation of *Accept*. This is due to, the missing attribute values of s_8 is considered similar to s_7.

In the following section, we present the proposed approaches called Similarity Class and an Enhanced Tolerance Relation.

III. SIMILARITY CLASS AND ENHANCED TOLERANCE RELATION

A. Similarity Class

From Definition 1, the concept of similarity class is defined as follows;

Definition 5. *Let S=(U,A,V,f), be a complete information system, and*
$\forall o_i,o_j\in U, Sim(o_i,o_j)\Leftrightarrow\forall a_k\in B,(a_k(o_i)=a_k(o_j))$,
the similarity class, $I_B^S(o_i)$ of an object *i*, with reference to set B, is defines as;

$$I_B^S(o_i)=\{o_j\mid o_j\in U\}$$

This can be elaborated using the following example. Table 2, is part of *test-cases* and *requirements* issues in software engineering, taken from [18], where o_i is a test-case *i*, and a_i is a requirement *i*.

Example 2: Table 2 is a complete *test-cases* and *requirements* information system, where $o_1,o_2,...,o_{10}$ are ten tests cases/ objects and $a_1,a_2,...,a_{10}$ are ten requirements/condition attributes, where their domain values are as shown in Table 2. The decision attribute, *d*, is with the value $\{pass, fail\}$,
$pass=\{o_2,o_3,o_4,o_6,o_7,o_8,o_9,o_{10}\}$ and $fail=\{o_1,o_5\}$.

From Definition 5, the results can be obtained by analyzing it with similarity class as follows:

TABLE 2: A COMPLETE INFORMATION SYSTEM FROM [18]

Obj's	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}	Decision (d)
o_1	1	1	1	1	1	1	0	0	0	0	*fail*
o_2	1	1	1	1	1	1	0	1	0	0	*pass*
o_3	1	1	1	1	1	1	0	0	0	0	*pass*
o_4	1	1	1	1	1	1	0	0	0	0	*pass*
o_5	1	1	1	1	1	1	0	0	0	0	*fail*
o_6	1	1	1	1	1	1	0	0	0	0	*pass*
o_7	1	0	1	1	1	0	1	0	0	0	*pass*
o_8	1	1	1	1	1	1	0	1	0	0	*pass*
o_9	1	1	1	1	1	1	0	0	1	0	*pass*
o_{10}	1	0	1	1	1	1	0	0	0	1	*pass*

$I_C^T(o_1)=\{o_1,o_3,o_4,o_5,o_6\}, I_C^T(o_2)=\{o_2,o_8\}$,
$I_C^T(o_3)=\{o_1,o_3,o_4,o_5,o_6\}$, $I_C^T(o_4)=\{o_1,o_3,o_4,o_5,o_6\}$
$I_C^T(o_5)=\{o_1,o_3,o_4,o_5,o_6\}, I_C^T(o_6)=\{o_1,o_3,o_4,o_5,o_6\}$,
$I_C^T(o_7)=\{o_7\}, I_C^T(o_8)=\{o_2,o_8\}, I_C^T(o_9)=\{o_9\}, I_C^T(o_{10})=\{o_{10}\}$

and

$I_{C\cup\{d\}}^T(o_1)=\{o_1,o_5\}, I_{C\cup\{d\}}^T(o_2)=\{o_2,o_8\}$,
$I_{C\cup\{d\}}^T(o_3)=\{o_3,o_4,o_6\}$, $I_{C\cup\{d\}}^T(o_4)=\{o_3,o_4,o_6\}$
$I_{C\cup\{d\}}^T(o_5)=\{o_1,o_5\}, I_{C\cup\{d\}}^T(o_6)=\{o_3,o_4,o_6\}$,
$I_{C\cup\{d\}}^T(o_7)=\{o_7\}, I_{C\cup\{d\}}^T(o_8)=\{o_2,o_8\}, I_{C\cup\{d\}}^T(o_9)=\{o_9\}$,
$I_{C\cup\{d\}}^T(o_{10})=\{o_{10}\}$

Thus,

$|S|=|\{\{o_1,o_5\},\{o_2,o_8\},\{o_3,o_4,o_6\},\{o_7\},\{o_9\},\{o_{10}\}\}|=6$

From the above result, there were indiscernible test cases in the similarity class, where o_1=o_5, o_2=o_8, and o_3=o_4=o_6. Thus, 10 objects can be reduced to 6 different objects

B. An Enhanced Tolerance Relation

The enhanced tolerance relation is basically based on the concept of similarity precision between objects x and y. However, the threshold value of similarity between objects x and y will be considered as discussed below.

Given an incomplete information system $S^B=(U,A,V_B,f)$, where $A=C\cup\{d\}$, C is a set of condition attributes and d the decision attribute, such that $f:U\times A\rightarrow V_a$. For any $a\in A$, where V_a is called domain of an attribute a and a subset $B\subseteq C$, the *similarity precision* is defined as follows.

Definition 6. *Let* $SP_B(x)=\{b\mid b\in B\wedge b(x)\neq *\}$, *the similarity precision* δ *, is defined as*

$$\delta(x,y)=\frac{|SP_B(x)\cap SP_B(y)|}{|C|}, \qquad (1)$$

From (1), it is clear that $0 < \delta(x,y) \le 1$. From Definition 6, the enhanced tolerance relation with similarity precision is given as follow:

Definition 7. Let an given IIS, $S^B = (U, A, V_B, f)$. The enhanced tolerance relation with similarity precision $L\delta$ is defined as follows;

$$\forall_{x,y\in U\times U} \quad (L\delta_B(x,y) \Leftrightarrow \forall_{b\in B}(b(x)=b(y)=*) \vee ((\delta(x,y)) \ge \alpha) \wedge \forall_{b\in B} \ (((b(x)\ne *) \wedge (b(y)\ne *)) \rightarrow (b(x)=b(y)))$$

where $\alpha \in (0,1]$ is a threshold value .

Since $\alpha \in (0,1]$, then $0<\delta(x,y)\le 1$ which implies that $SP_B(x) \cap SP_B(y) \ne \phi$ holds, but not vice versa if certain threshold value of the similarity is given.

To clearly depict the enhanced tolerance class as defined above, we illustrate through an example as follows.

Example 3. Two objects $x = \{0,1,*,*,*,*,0,0\}$ and $y = \{0,1,*,0,0,0,*,*\}$ are tolerant if it is based on enhanced tolerance relation, i.e., $SP(x) \cap SP(y) = \{1,0\}$ or $\delta(x,y) = 0.25$.

From the value of $\delta(x,y)$, we believed that both objects are less tolerant. Moreover, if we set $\alpha = 0.5$, then $(x,y) \notin L\delta$ .That is the two objects are not tolerant if the similarity precision does not hold the threshold value.

The above relation is reflexive and symmetric but not necessarily transitive. To clearly depict the enhanced tolerance with similarity precision as defined above, we illustrate through an example from Table 1.

Example 4. From Table 1, two objects s_1 and s_8 are not tolerant if $\alpha = 0.4$. However, two objects s_4 and s_5 are tolerant due to $\delta(a_4, a_5) \ge 0.5$.

In the following sub-section, we present two properties for enhanced tolerance relation and their correctness proofs.

C. *Correctness of Proof*

Proposition 1. *Let given an IIS,* $S^B = (U, A, V_B, f)$, *a subset* $B \subseteq C$ *and* $x \in U$. *If* $\delta > 0$, *then*

a. *For any x and y,* $L\delta_B(x,y) \Rightarrow L_B(x,y)$

b. $L\delta_B(x,y) \Leftarrow L_B(x,y)$ *except when* $SP_B(x) \cap SP_B(y) = \phi$.

Proof

a. When $\delta > 0$, then
$$L\delta_B(x,y) \Leftrightarrow \alpha_B(x,y) > 0$$
$$\Leftrightarrow SP_B(x) \cap SP_B(y) \ne \phi$$
$$\wedge \forall a \in SP_B(x) \cap SP_B(y), f_a(x) = f_a(y)$$
$$\Rightarrow L_B(x,y)$$

b. It is clear that $L_B(x,y) \Rightarrow L\delta_B(x,y)$ except when $SP_B(x) \cap SP_B(y) = \phi$.

Definition 8. *Let an given IIS,* $S^B = (U, A, V_B, f)$ *and* $B \subseteq C$. *The new enhanced tolerance class is defined as* $I_B^{L\delta}(x) = \{y \mid y \in U \wedge L\delta_B(x,y)\}$.

To clearly depict the enhanced tolerance class as defined above, we illustrate through an example from Table 1.

Example 5. From Table 1, and let $\delta \ge 0.5$, we have an enhanced tolerance classes as follows

$I_C^{L\delta}(s_1) = \{s_1, s_4, s_5, s_9\}$, $I_C^{L\delta}(s_2) = \{s_2, s_5, s_6, s_7\}$, $I_C^{L\delta}(s_3) = \{s_3\}$,
$I_C^{L\delta}(s_4) = \{s_1, s_4, s_5, s_9\}$, $I_C^{L\delta}(s_5) = \{s_1, s_2, s_4, s_5, s_6, s_7, s_9\}$,
$I_C^{L\delta}(s_6) = \{s_2, s_5, s_6, s_7\}$, $I_C^{L\delta}(s_7) = \{s_2, s_5, s_6, s_7\}$,
$I_C^{L\delta}(s_8) = \{s_8\}$, $I_C^{L\delta}(s_9) = \{s_1, s_4, s_5, s_9\}$

and

$Accept_C^{L\delta} = \{s_1, s_2, s_4, s_5, s_6, s_7, s_9\}$,
$\mathrm{Re}\,ject_C^{L\delta} = \{s_3, s_8\}$,

$Accept_{L\delta}^C = \{s_1, s_2, s_4, s_5, s_6, s_7, s_9\}$,
$\mathrm{Re}\,ject_{L\delta}^C = \{s_3, s_8\}$

From the above analysis, the results of the proposed approach are more flexible and precise as compared to the previous approaches [20, 22], where in this case s_7 and s_8 are divided into different class.

Definition 9. *Let an given IIS,* $S^B = (U, A, V_B, f)$. *The lower approximation and the upper approximation of an object x based on an enhanced tolerance class* $I_B^{L\delta}(x)$ *denoted as* $ET_{L\delta}^B(x)$ *and* $ET_B^{L\delta}(x)$ *respectively are defined as*

$$ET_B^{L\delta} = \{x \mid x \in U \wedge I_B^{L\delta}(x) \subseteq ET\}, \text{ and}$$
$$ET_{L\delta}^B = \{x \mid x \in U \wedge I_B^{L\delta}(x) \cap T \ne \phi\}.$$

From Definition 9, we can generalize Proposition 1 as described in the following proposition.

Proposition 2. *Let given an incomplete information system* $S^* = (U, A, V_*, f)$, *a subset* $B \subseteq A$ *and* $x \in U$. *If* $0 \le \delta_1 < \delta_2 \le 1$, *then* $I_B^{L\delta_2} \subseteq I_B^{L\delta_1}$.

Proof

For every $a \in I_B^{L\delta_2}(x)$, we have $\alpha_B(x,y) \ge \delta_2$. Since $\delta_2 > \delta_1$, then $\alpha_B(x,y) \ge \delta_1$, that is $\forall a \in I_B^{L\delta_1}(x)$ which implies $I_B^{L\delta_2}(x) = I_B^{L\delta_1}(x)$. However, if $\alpha_B(x,y) \ge \delta_1$ then it does not necessarily $\alpha_B(x,y) \ge \delta_2$. Hence $I_B^{L\delta_2} \subseteq I_B^{L\delta_1}$.

To clearly depict the property of generalized enhanced tolerance class in Proposition 2, we illustrate through an example from Table 1.

Example 6. From Table 1, we have $I_C^{L\delta_1}(s_2) = I_C^{L\delta_2}(s_6) = \{s_2, s_5, s_6, s_7\}$ for $\delta_1 = 0.5$. However, for $\delta_2 = 0.75$, we have $I_C^{L\delta_2}(s_6) = \{s_6, s_7\}$ and thus, $I_C^{L\delta_2}(s_6) \neq I_C^{L\delta_1}(s_6)$.

Definition 10. From definition 7, the reduction of $L\delta$ is given as follows;
$R^{L\delta}(x) = \{x \mid x \in L\delta_B(x, y) \wedge |b(x)| > |b(y)|, b(x) and b(y) \neq '*'\}$,
and the reduction of U can be defined as;

$$R^U = |\{x \in U : R^{L\delta}(x)\}|$$

Thus, from the above example,
$|L\delta| = |\{s_1, s_4, s_9\}, \{s_2, s_6, s_7\}, \{s_3\}, \{s_8\}| = 4$

From the above result, there were indiscernible students from enhanced tolerance relation, where $s_1 = s_4 = s_9$ and $s_2 = s_6 = s_7$. As such, the nine different students could easily be reduced to four different classification of students when $\delta = 0.5$.

IV. EXPERIMENTAL RESULTS

This section presents the number of data reductions as compared to other approaches. The comparisons were executed on PC with 2.53GHz CPU, 4.0GB RAM and Windows 7 Professional. In evaluating the effectiveness of the reduced size of the reduction rate will be used for complete information systems, while for incomplete information systems, the reduction rate and the accuracy are considered.

Definition 10: *Let m(o) be the number of reduced data objects, n(o) be the number of actual data objects. The reduction rate is calculated as;*

$$\text{Reduction Rate} = \frac{m(o)}{n(o)} x100\ \%$$

Definition 11. . *Let* $S^B = (U, A, V_B, f)$ *be an IIS, and* $B \subseteq A$. *The lower approximation is* $x_B^T = \{x \mid x \in U \wedge I_B^T(x) \subseteq X\}$ *and the upper approximation is* $\bar{x}_B^T = \{x \mid x \in U \wedge I_B^T(x) \cap X \neq \phi\}$. *The accuracy of approximation of an object set X with reference to attribute set B can be defined as*

$$Accuracy = I_B^T(x) / I_T^B(x)$$

Table 3 shows that, Gotlieb *et al.* [23] has the lowest reduction rate as compared to Xu et.al. [24] and the proposed approach. However the proposed approach has better with 40% reduction rate which subsequently require lower execution time.

The Soybean information system can be seen as in appendix 1. The same procedure is used as what we have analyzed from the above scholarship-application-student incomplete information system. From Table 4, the reduction rate is similar

TABLE 3: A COMPARISON BETWEEN PROPOSED APPROACH WITH TWO OTHER APPROACHES FOR COMPLETE INFORMATION SYSTEM.

authors	Technique	#actual obj's	#actual dimen's	#obj. after reduc.	Reduc rate
Gotlieb et al. [23]	FLOWER	10	10	8	20%
Xu et. al. [24]	Weighted Greedy Algorithm	10	10	7	30%
Proposed approach	Similarity class	10	10	6	40%

TABLE 4: A COMPARISON BETWEEN PROPOSED APPROACH WITH TOLERANCE RELATION FOR INCOMPLETE INFORMATION SYSTEM (SOYBEAN).

Tech.	#actual obj.s	#actual dimen's	#obj after reduc.	Reduc. rate	Accuracy
Tolerance relation	20	6	13	35%	30%
Enhanced tolerance relation	20	6	13	35%	70%

to both approaches, however, the proposed approach (enhanced tolerance relation) outperformed in terms of accuracy.

V. CONCLUSION

Similarity Class was used to reduce the objects for structured and complete information systems, while the enhanced tolerance relation was used to reduce the structured and incomplete information system. The result showed that, the reduced objects using similarity class produced a greater reduction rate and subsequently, reduced execution time and space as compared with the actual data. For the case of structured and incomplete information systems, the enhanced tolerance relation showed better results in terms of accuracy as compared with toleration relation approach.

Further research will consider the huge volume of data size/ information system as well as taking into consideration on the dimensions reduction which is one of the main issues in big data analytics.

ACKNOWLEDGEMENT

The research was supported from Ministry of Higher Education through Fundamental Research Grant Scheme (FRGS) vote number 1643 and special grant scheme from UTHM with vote number U689.

REFERENCES

[1] M.D. Mustafa, N.F. Nabila, D.J. Evans, M.Y. Saman, A. Mamat, "Association rules on significant rare data using second support",

International Journal of Computer Mathematics 83 (1), 69-80, 2006
[2] C.S Greene, J. Tan, M. Ung, J.H. Moore, C.Cheng," Big Data Bioinformatics", *J. Cell Physiol*,229(12), 1896-900, 2014
[3] A. L'Heureux, K. Grolinger,H.F. Elyamany, M.A.M. Capretz,"Machine Learning with Big Data: Challenges and Approaches", *IEEE Access* vol.5, 7776-7797, 2017
[4] S. Trabelsi, Z. Elouedi, and P. Lingras,"Classification systems based on rough sets under the belief function network", *International Journal of Approximate Reasoning*, vol.52 (9), pp.1409-1432, 2011.
[5] K. Kaneiwa,"A rough set approach to multiple dataset analysis", *Journal of Applied Soft Computing*, vol. 11 (2), pp. 2538-2547, 2011
[6] M.H. ur Rehman, C.S. Liew, A. Abbas, P.P. Jayaraman, T.Y. Wah, S.U. Khan,"Big Data Reduction Methods: A Survey", *Data Sci. Eng.* 1, 265-284, 2016
[7] Z. Pawlak,"Rough sets", *International Journal of Computer and Information Sciences*, vol. 11 (5), pp. 341-356, 1982.
[8] Z. Lu and Z. Qin," 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, 2011.
[9] J. Dai, W. Wang, Q. Xu, and H. Tian," Uncertainty measurement for interval-valued decision systems based on extended conditional entropy", *Knowledge-based Systems*, vol. 27, pp. 443-450, 2012.
[10] A. Skowron and P. Wasilewski," Toward interactive Rough-Granular Computing", *Control and Cybernetics*, vol. 40, no. 2, pp. 213-235, 2011.
[11] A. Skowron, J. Stepaniuk, and R. Swiniarski," Approximation spaces in Rough-Granular Computing", *Fundamentae Informaticae*, vol. 100, no. 1-4, pp. 141-157, 2010.
[12] I.T.R.Yanto, P. Vitasari, T.Herawan, and M.M.Deris,"Applying variable precision rough set model for clustering suffering student's enxiety", *Expert Systems with Applications*, vol. 39(1), 452-459, 2012.
[13] T.Herawan, M.M.Deris, and J.H.Abawajy,"A rough set approach for selecting clustering attributes", *Knowledge-Based Systems*, vol. 23(3), pp.220-231, 2010
[14] D. Parmar, T. Wu, and J. Blackhurst,"MMR: An algorithm for clustering categorical data using rough set theory", *Data & Knowledge Engineering*, vol. 63(3), pp. 879-893, 2007.
[15] D. Kim," Data classification based on tolerant rough set", *Pattern Recognition*, vol. 34(8), pp. 1613-1624, 2001.
[16] T. Yan and C. Han,"A novel approach of rough conditional entropy-based attribute selection for incomplete decision system", *Mathematical Problems in Engineering*, vol. 2014, pp.1-15, 2014
[17] J.W. Grzymala-Busse,"Rough set strategies to data with missing attribute values" *Proceeding of the workshop on Foundation and New Directions in Data Mining, associated with the 3rdIEEE International Conference on Data Mining, pp. 56-63, 2003*
[18] N.F.M. Nasir, N Ibrahim, MM Deris, MZ Saringat," Test case and requirement selection using rough set theory and conditional entropy",Computational Intelligence in Information Systems", Springer, 61-71, 2018
[19] J. Stefanowski and A. Tsoukias,"Incomplete information table and rough classification", *Computational Intelligence*, vol. 17 (3), pp. 545-566, 2001.
[20] Q Zhou," Research on Tolerance-based Rough Set Models", IEEE conf. on System Science, Engineering Design, Manufacturing Information, 137-139,2010
[21] M.M. Deris, Z.Abdullah, R. Mamat, Y.Yuan,"A New Limited Tolerance Relation for Attribute Selection in Incomplete Information Systems", IEEE Conf. on Fuzzy Systems and Knowledge Discovery (FSKD), 964-970, 2015
[22] H. Huang, G. Wang, and Y. Wu,"A Direct approach for incomplete information systems", *Data Mining and Knowledge Discovery: Theory ,Tools and Technology*, VI, 114, 2004
[23] A. Gotlieb & D. Marijan ," *FLOWER: optimal test suite reduction as a network maximum flow",* International Symposium on Software Testing and Analysis (ISSTA),171-180, 2014.
[24] S. Xu, H. Miao & H. Gao," *Test suite reduction using weighted set covering techniquee",* Software Engineering, Artificial Intelligence, Networking and Parallel & Distributed Computing (SNPD), 307-312, 2012

Appendix 1

Soybean	Stem	Seed-discolor	Seed-size	Shriveling	Roots	Mycelium	Decision
1	Abnorm	*	Norm	Absent	Normal	Absent	Present
2	Norm	Absent	Norm	Absent	Normal	Absent	Present
3	Abnorm	Present	Norm	*	Normal	Absent	Present
4	Abnorm	Absent	Norm	Absent	Galls-cysts	Present	Present
5	*	Present	It-Norm	Absent	Normal	Absent	Present
6	Norm	Absent	Norm	*	Normal	*	Present
7	Abnorm	Absent	Norm	Absent	Normal	Absent	Present
8	Abnorm	Absent	*	Absent	Galls-cysts	Absent	Present
9	*	Absent	Norm	Absent	Normal	Absent	Present
10	Abnorm	Absent	Norm	Absent	Normal	Absent	Present
11	Abnorm	*	It-Norm	*	Galls-cysts	*	Present
12	Norm	*	Norm	*	Galls-cysts	*	Present
13	Norm	*	It-Norm	*	Normal	Absent	Present
14	Norm	*	It-Norm	*	Galls-cysts	*	Absent
15	Abnorm	Present	It-Norm	*	Galls-cysts	*	Absent
16	Norm	*	It-Norm	*	Galls-cysts	Present	Absent
17	Norm	*	*	*	*	*	Absent
18	*	*	*	*	Rotted	*	Present
19	Abnorm	*	Norm	*	Rotted	Present	Absent
20	Abnorm	*	*	*	Rotted	*	Present

arity_Class_and_Enhanced_Tolerance_Relation_for_Complete...

ORIGINALITY REPORT

14%	%	14%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	"Recent Advances on Soft Computing and Data Mining", Springer Science and Business Media LLC, 2018 Publication	9%
2	"Computational Intelligence in Information Systems", Springer Science and Business Media LLC, 2019 Publication	3%
3	"Recent Advances on Soft Computing and Data Mining", Springer Nature, 2017 Publication	2%
4	Renxia Wan, Duoqian Miao, Witold Pedrycz. "Constrained tolerance rough set in incomplete information systems", CAAI Transactions on Intelligence Technology, 2021 Publication	2%

Exclude quotes On
Exclude bibliography On
Exclude matches < 2%