Journal of Mathematics Research; Vol. 8, No. 6; December 2016
ISSN 1916-9795 E-ISSN 1916-9809
Published by Canadian Center of Science and Education

Discovery of Similarity and Dissimilarity

T. N. Alharthi[1] & M. A. Elsafty[2]

[1] Departement of Mathematics, Faculty of Science, Bisha University, Saudi Arabia

[2] Departement of Mathematics, Faculty of Science, Taif University, Saudi Arabia

Correspondence: T. N. Alharthi, Departement of Mathematics, Faculty of Science, Bisha University, Saudi Arabia. E-mail: thoraya-alharthi@hotmail.com

Received: September 11, 2016 Accepted: October 8, 2016 Online Published: November 25, 2016

doi:10.5539/jmr.v8n6p105 URL: http://dx.doi.org/10.5539/jmr.v8n6p105

Abstract

The aim of this paper is to introduce some applications on similarities and dissimilarities. Using of a simplified diagram and tables to present the information about the similarities and dissimilarities account process and organization are also easy and we calculated the topology based on the similarity and topology views of the dissimilarity. For information system whose values are numeric, a method of classification is suggested. This method is based on constructing neighborhood relation on the universe of the resulted classification not generally a partition for the universe.

Keywords: rough sets, topology, similarity relation

1. Introduction

Similarity is necessary for knowledge discovery. Granulation, classification, and cluster analysis each include some notion or a definition of similarity. The domain and distribution of the data are the base measurement of similarity were selected. Some similarity metrics may be considered more of use than others even within a specific domain. There is some uncertainty in quantitative measurement of similarity between records of mixed data. This uncertainty comes from the lack of scale that nominal and ordinal data have. Rough set theory is a tool that is developed for the sake of handling uncertainty. Rough sets may be used in dissimilarity analysis of qualitatively-collected data. It would seem that rough sets can be used in measuring similarity between records which contain quantitative and qualitative data for clustering the records. Rough sets were considered one of the tools that have been developed to deal with uncertainty. Rough sets measure of similarity between our records that contain the data with same qualitative and quantitative data (Han and Kamber, 2001; Lin et al., 2002; Pawlak, 1982; Pawlak, 2002; Zhu, 2002).

The knowledge can be express by Mathematics, whether the knowledge contains quantitative or qualitative.

Data types

1) **Quantitative data** is information regarding quantities, which is information that could be measured and written using numbers.

For example:

- Student's grades in school materials.

- Degree heat of patients in the hospital.

2) **Qualitative data** can be described as ordinal or nominal. Nominal data does not have order nor scale.

For example:

- Cities.

Ordinal data has order without scale. For example:

- Colors.

2. Similarity and Dissimilarity in Rough Sets (Han and Kamber, 2001)

Pawlak gives description for applying rough sets to measurement of dissimilarity between records of Boolean values.

Example 1.1

Let marks of 4- students, $C= \{c_1, c_2, c_3, c_4\}$ and their subjects, $A=\{ a_1, \ldots, a_7\}$, where, a_1= Mathematics, a_2= History, a_3=Geography, a_4 =Science, a_5= English, a_6= Physics and a_7 = Chemistry.

Table 1.1

	a_1	a_2	a_3	a_4	a_5	a_6	a_7
c_1	88	100	100	100	90	90	88
c_2	88	90	90	90	100	100	88
c_3	88	90	90	90	90	100	88
c_4	88	90	100	90	100	100	88

We choose attributes a_2, a_3 and a_5 from the Table 1.1, we get **Table 1.2**

Possible modified set Table 1.2

	a_2	**a_3**	a_5
c_1	100	100	90
c_2	90	90	100
c_3	90	90	90
c_4	90	100	100

∗Any attributes with the same value as another attribute for all records are disregarded.

Each record has a node and a label edge between the nodes if deleting an attribute would place the records in the same class of equivalence.

For example: degree is between $c2$ and $c4$ with the label $a5$. We get the result as the next figure as follow:

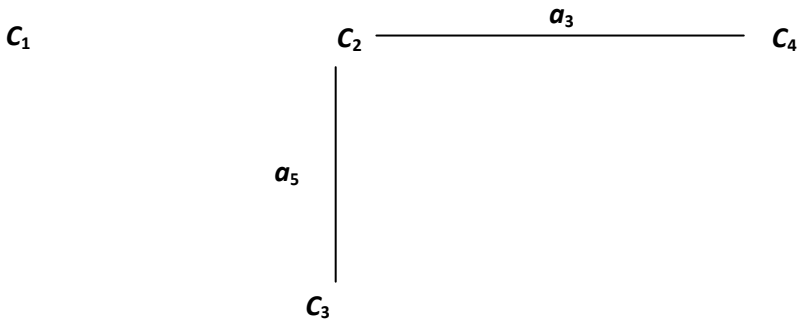

Figure 1.1

Determining length of the shortest path between the nodes in the graph corresponding to the records is the way dissimilarity between two records is computed. For example:

The dissimilarity between c_3 and c_4 would be 2.

Definition 1.1

The similarities are computed as: $(|D_{max}-D_{ij})/D_{max}|$, where D_{max} is the maximum dissimilarity over all pairs and D_{ij} is the dissimilarity between c_i and c_j, D_{max} is the previous example is 2.

The dissimilarity is computed out of (through) the following:

Discernibility Matrices:

Definition 1.2

An information system S defines a matrix M_A, which is called discernibility matrix. Each entry $M_A(x, y) \subseteq A$ consists of a set of attributes that be used to discern between objects $x, y \in U$:

M_A is a $|U|\times|U|$ matrix; the discernibility matrix has the form:

$$M_{ij} = \{a \in A: a(x_i) \neq a(x_j)\} \text{ for } i, j \in \{1, 2, \ldots, n\}, n=| \quad |.$$

Table 1.3 pair wise similarities form Table 1.2

	c_1	c_2	c_3	c_4
c_1	1	1	1	1
c_2	1	1	0.5	0.5
c_3	1	0.5	1	0
c_4	1	0.5	0	1

Where $c_{11} = (|D_{max}-D_{ij})/D_{max}|) = c_{12} = c_{13}=c_{14} =2-0/2=1$

Table 1.4 Discerniblity matrix of Table 1.1

	$\boldsymbol{c_1}$	$\boldsymbol{c_2}$	$\boldsymbol{c_3}$	c_4
c_1	Ø	$\{a_2, a_3, a_4, a_5, a_6\}$	$\{a_2, a_3, a_4, a_6\}$	$\{a_2, , a_4, a_5, a_6\}$
c_2	$\{a_2, a_3, a_4, a_5, a_6\}$	Ø	$\{a_5\}$	$\{a_3\}$
c_3	$\{a_2, a_3, a_4, a_6\}$	$\{a_5\}$	Ø	$\{a_3, a_5\}$
c_4	$\{a_2, a_4, a_5, a_6\}$	$\{a_3\}$	$\{a_3, a_5\}$	Ø

Table 1.5 dissimilarity

	$\boldsymbol{c_1}$	$\boldsymbol{c_2}$	$\boldsymbol{c_3}$	c_4
c_1	0	5	4	4
c_2	5	0	1	1
$c3$	4	1	0	2
c_4	4	1	2	0

Example 1.2

Let $U=\{c_1, c_2, c_3, c_4, c_5, c_6, c_7, c_8\}$ be a mobile devices, $A=\{ a_1, a_2, a_3\}$ be screen measurement, weight and accuracy of camera in the following Table 1.6

Table 1.6

U/A	$\boldsymbol{a_1}$	$\boldsymbol{a_2}$	a_3
c_1	5	116	8
c_2	4.8	130	13
c_3	4.8	130	8
c_4	5	133	13
c_5	4.8	116	13
c_6	4.3	133	13
c_7	4.3	116	8
c_8	5	116	13

We draw the relation between the objects and attributes as follows

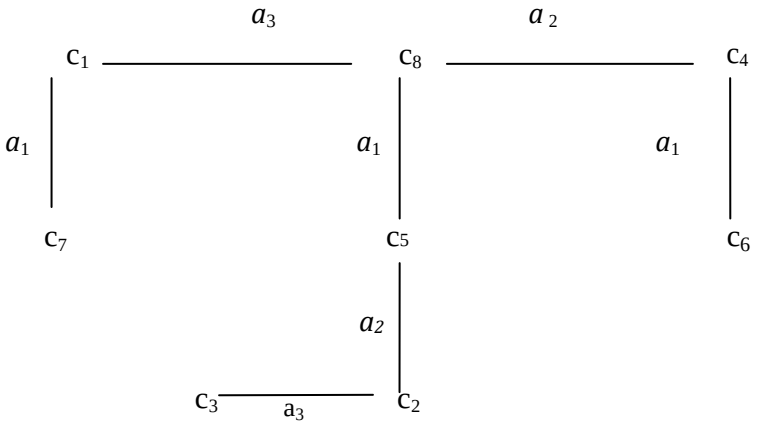

Figure 1.2. Pair wise similarities for Table 1.6

Dmax in this example is 5 where, c_{11}=5−0/5=1, c_{12}=c_{16}=5−3/5=2/5, c_{13}=5−4/5=1/5, c_{14}=c_{15}=5−2/5=3/5, c_{17}=c_{18}=5−1/5=4/5.

Table 1.7

	c_1	**c_2**	**c_3**	**c_4**	**c_5**	**c_6**	**c_7**	c_8
c_1	1	2/5	1/5	3/5	3/5	2/5	4/5	4/5
c_2	2/5	1	4/5	2/5	4/5	1/5	1/5	3/5
c_3	1/5	4/5	1	1/5	3/5	0	0	2/5
c_4	3/5	2/5	1/5	1	3/5	4/5	2/5	4/5
c_5	3/5	4/5	3/5	3/5	1	2/5	2/5	4/5
c_6	2/5	1/5	0	4/5	2/5	1	1/5	3/5
c_7	4/5	1/5	0	2/5	2/5	1/5	1	3/5
c_8	4/5	3/5	2/5	4/5	4/5	3/5	3/5	1

Table 1.8 Discernibility matrix of Table 1.6

	c_1	**c_2**	**c_3**	**c_4**	**c_5**	**c_6**	**c_7**	c_8
c_1	Ø	{a_2, a_3, a_3}	{a_1, a_2}	{a_2, a_3}	{a_1, a_3}	{a_2, a_3, a_3}	{a_1}	{a_3}
c_2	{a_2, a_3, a_3}	Ø	{a_3}	{a_1, a_2}	{a_2}	{a_1, a_2}	{a_2, a_3, a_3}	**{a_1, a_2}**
c_3	{a_1, a_2}	{a_3}	Ø	{a_2, a_3, a_3}	{a_2, a_3}	{a_2, a_3, a_3}	{a_1, a_2}	**{a_2, a_3, a_3}**
c_4	{a_2, a_3}	{a_1, a_2}	{a_2, a_3, a_3}	Ø	{a_1, a_2}	{a_1}	{a_2, a_3, a_3}	{a_2}
c_5	{a_1, a_3}	{a_2}	{a_2, a_3}	{a_1, a_2}	Ø	{a_1, a_2}	{a_1, a_3}	{a_1}
c_6	{a_2, a_3, a_3}	{a_1, a_2}	{a_2, a_3, a_3}	{a_1}	{a_1, a_2}	Ø	{a_2, a_3}	**{a_1, a_2}**
c_7	{a_1}	{a_2, a_3, a_3}	{a_1, a_2}	{a_2, a_3, a_3}	{a_1, a_3}	{a_2, a_3}	Ø	**{a_2, a_3, a_3}**
c_8	**{a_3}**	{a_1, a_2}	{a_2, a_3, a_3}	**{a_2}**	**{a_1}**	{a_1, a_2}	{a_2, a_3, a_3}	Ø

Table 1.9 Dissimilarity

	c_1	**c_2**	**c_3**	**c_4**	**c_5**	**c_6**	**c_7**	c_8
c_1	0	3	2	2	2	3	1	1
c_2	3	0	1	2	1	2	3	2
c_3	2	1	0	3	2	3	2	3
c_4	2	2	3	0	2	1	3	1
c_5	2	1	2	2	0	2	2	1
c_6	3	2	3	1	2	0	2	2
c_7	1	3	2	3	2	2	0	3
c_8	1	2	3	1	1	2	3	0

To compute Similarity matrix of Table 1.6 in the Table 1.10 with attribute a1 as follows
$|c(x_i)\text{-}c(y_j)|$, $\forall i, j \in \{1,2,3,4,5,6,7,8\}$, where
$c_{11}=x_{14}= x_{18}=|5\text{-}5|=0$, $c_{12}= c_{13}= c_{15}=|5\text{-}4.8|=0.2$,
$c_{16}= c_{17}=|5\text{-}4.3|=0.7$.

Table 1.10

a_1	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8
c_1	0	0.2	0.2	0	0.2	0.7	0.7	0
c_2	0.2	0	0	0.2	0	0.5	0.5	0.2
c_3	0.2	0	0	0.2	0	0.5	0.5	0.2
c_4	0	0.2	0.2	0	0.2	0.7	0.7	0
c_5	0.2	0	0	0.2	0	0.5	0.5	0.2
c_6	0.7	0.7	0.7	0.7	0.7	0	0	0.7
c_7	0.7	0.7	0.7	0.7	0.7	0	0	0.7
c_8	0	0	0	0	0	0.7	0.7	0

To compute the Table 1.11 with attribute a_2 as follows, where $c_{11} = c_{15} = c_{17} = c_{18}$ $= | 116 – 116 | = 0$, $c_{12} = c_{13} = | 116 – 130 | = 14$, $c_{14} = c_{16} = |116 – 133 | = 17$.

Table 1.11

a_2	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8
c_1	0	14	14	17	0	17	0	0
c_2	14	0	0	3	14	3	14	14
c_3	14	0	0	3	14	3	14	14
c_4	17	3	3	0	17	0	17	17
c_5	0	14	14	17	0	17	0	0
c_6	17	3	3	0	17	0	17	17
c_7	0	14	14	17	0	17	0	0
c_8	0	14	14	17	0	17	0	0

The similarity values of Table 1.12 is compute as the fallow, where $x_{11} = x_{13} = x_{17} = |5-5| = 0$, $x_{12} = x_{14} = x_{15} = x_{16} = x_{18} = |8-13| = 5$.

Table 1.12

a_3	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8
c_1	0	5	0	5	5	5	0	5
c_2	5	0	5	0	0	0	5	0
c_3	0	5	0	5	5	5	0	5
c_4	5	0	5	0	0	0	5	0
c_5	5	0	5	0	0	0	5	0
c_6	5	0	5	0	0	0	5	0
c_7	0	5	0	5	5	5	0	5
c_8	5	0	5	0	0	0	5	0

Example 1.3

Let $U=\{x_1, x_2, x_3, x_4\}$ be patients, $A=\{a_1, a_2, a_3\}$ be Temperature, pressure and Diabetics in the following table:

Table 1.13

	a_1	a_2	a_3
x_1	37	100	70
x_2	40	120	70
x_3	37	120	70
x_4	45	100	70

We graph the relation between attributes and objects as the next figure

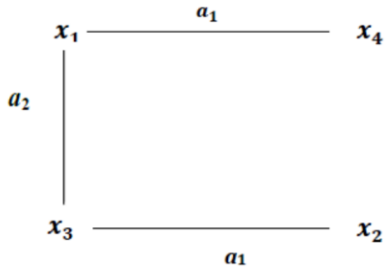

Figure 1.3

To compute similarities for Table 1.14, where $D_{\max} = 3$ as follows

Table 1.14

	x_1	x_2	x_3	x_4
x_1	1	1/3	2/3	2/3
x_2	1/3	1	2/3	0
x_3	2/3	2/3	1	1/3
x_4	2/3	0	1/3	1

Definition 1.3

If (U, A, V, f) is an information system defines an information function $f: U \rightarrow V$, where A is the set of attributes, V is the domain of the particular attributes in which the values V are real numbers. We define a relation R_i for each objects $i(x)$ as follows: $xR_i\ y$ if $|i(x)-i(y)|<\lambda$, where λ is determined by an expert of the field. For example if the information is from medical field, the expert is a person interested in medicine and making in the problem. Thus for each $i(x) \in U$ we can get a classification O/R_i where O is a finite set, which is $xR_i=\{y:|i(x)-i(y)|<\varepsilon, x \in 0\}$.

Definition 1.4

For each B $\subset$ A, the relation $R_B \subset U \times U$ defined $x\ R_B\ y = \frac{\sum_{i \in B} |i(x)- i(y)|}{|B|} < \lambda$, where $|B|$ is the cardinality of B and λ is a represented any number.

Yao's method (Yao, 1999)

Yao introduced a method for generalization of approximation space depending on the right neighborhood as showing:

If U is a finite universe and R is a binary relation on U, then:

The class of right neighborhood is $(x)_R = \{y \in U: x\ R\ y\}$. For a topological space (X, t), a subset A of X, we define the accuracy of Yao is $\left|A^o/_{\bar{A}}\right|$.

Choose $\lambda \leq \frac{1}{3}$ is the ratio of specific similarity of the Table 1.14 and us finding the subbase information system as follows:

$x\ R\ y = \{(x_1, x_2), (x_2, x_1), (x_2, x_4), (x_3, x_4), (x_4, x_2), (x_4, x_3)\}$, then

$x_1R=\{x_2\}$, $x_2R=\{x_1, x_4\}$, $x_3R=\{x_4\}$, $x_4R=\{x_2, x_3\}$

$(x)_R$={{x_2}, {x_1, x_4}, {x_4}, {x_2, x_3}} is a class right neighborhood as in Yao's method.

S_R={{x_2}, {x_1, x_4}, {x_4}, {x_2, x_3}} is a subbase of τ_R as in TAS's method.

In TAS's method "Topological Approximation Space", we get:

B_R={Ø, {x_2}, { x_4}, {x_1, x_4}, {x_2, x_3}},

τ_R={Ø, X, {x_2}, {x_4}, {x_1, x_4}, {x_2, x_3}, {x_2, x_4}, {x_1, x_2, x_4},{x_2, x_3, x_4}} and

={Ø, X, {x_1, x_3, x_4}, {x_1, x_2, x_3}, {x_2, x_3}, {x_1, x_4}, {x_1, x_3}, {x_3},{x_1}}.

We find lower and upper approximation, closure and interior for all subset of U (2^4=16 subset) to get the accuracy by using Yao's method and TAS's method as shown in Table 1.15.

Table 1.15

P(X)	P(X^C)	**Yao's method**		**accuracy**	**TAS's method**		**accuracy**
		$X_{int(x)}$	$X_{cl(x)}$		$\underline{X}$	$\overline{X}$	
Ø	***X***	**Ø**	**Ø**	**0**	**Ø**	**Ø**	**0**
{x_1}	**{x_2, x_3, x_4}**	**Ø**	**{x_1}**	**0**	**Ø**	**{x_1, x_4}**	**0**
{x_2}	**{x_1, x_3, x_4}**	**{x_2}**	**{x_2, x_3}**	**1/2**	**{x_2}**	**{x_2, x_3}**	**1/2**
{x_3}	**{x_1, x_2, x_4}**	**Ø**	**{x_3}**	**0**	**Ø**	**{x_2, x_3}**	**0**
{x_4}	**{x_1, x_2, x_3}**	**{x_4}**	**{x_1, x_4}**	**1/2**	**{x_4}**	**{x_1, x_4}**	**1/2**
{x_1, x_2}	**{x_3, x_4}**	**{x_2}**	**{x_1, x_2, x_3}**	**1/3**	**{x_2}**	**{x_1, x_2, x_4}**	**1/3**
{x_1, x_3}	**{x_2, x_4}**	**Ø**	**{x_1, x_3}**	**0**	**Ø**	***X***	**0**
{x_1, x_4}	**{x_2, x_3}**	**{x_1, x_4}**	**{x_1, x_4}**	**1**	**{x_1, x_4}**	**{x_1, x_4}**	**1**
{x_2, x_3}	**{x_1, x_4}**	**{x_2, x_3}**	**{x_1, x_3}**	**1**	**{x_2, x_3}**	**{x_2, x_3}**	**1**
{x_2, x_4}	**{x_1, x_3}**	**{x_2, x_4}**	***X***	**1/2**	**{x_2, x_4}**	***X***	**1/2**
{x_3, x_4}	**{x_1, x_2}**	**{x_4}**	**{x_1, x_3, x_4}**	**1/3**	**{x_4}**	***X***	**1/3**
{x_1, x_2, x_3}	**{x_4}**	**{x_2, x_3}**	**{x_1, x_2, x_3}**	**2/3**	**{x_2, x_3}**	***X***	**2/3**
{x_1, x_2, x_4}	**{x_3}**	**{x_1, x_2, x_4}**	***X***	**3/4**	**{x_1, x_2, x_4}**	***X***	**3/4**
{x_2, x_3, x_4}	**{x_1}**	**{x_2, x_3, x_4}**	***X***	**3/4**	**{x_2, x_3, x_4}**	***X***	**3/4**
{x_1, x_3, x_4}	**{x_2}**	**{x_1, x_4}**	**{x_1, x_3, x_4}**	**2/3**	**{x_1, x_4}**	***X***	**1/2**
X	**Ø**	***X***	***X***	**1**	***X***	***X***	**1**

3. Fusing Quantitative and Qualitative Information (Han and Kamber, 2001)

Methods and metrics have become cluster data records that have quantitative or qualitative data only. It is possible to extract information by the fusion of the methods result or the various measures.

Quantitative measures depend on the complete continuous interval while qualitative measures depend on a discrete linear subset of the interval [0, 1].

Example 1.4

Let U={c_1, c_2, c_3, c_4, c_5, c_6, c_7, c_8} be a mobile devices, A={ a_1, a_2, a_3, a_4} be screen measurement, weight, accuracy of camera and colors in the following

Table 1.16

	a_1	a_2	a_3	a_4
c_1	5	116	8	*B*
c_2	4.8	130	13	*W*
c_3	4.8	130	8	*B*
c_4	5	133	13	*S*
c_5	4.8	116	13	*S*
c_6	4.3	133	13	*W*
c_7	4.3	116	8	*W*
c_8	5	116	13	B

Table 1.17

a_4	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8
c_1	0	1	0	1	1	1	1	0
c_2	1	0	1	1	1	0	0	1
c_3	0	1	0	1	1	1	1	0
c_4	1	1	1	0	0	1	1	1
c_5	1	1	1	0	0	1	1	1
c_6	1	0	1	1	1	0	0	1
c_7	1	0	1	1	1	0	0	1
c_8	0	1	0	1	1	1	1	0

Using Table 1.17and 1.12 to find Table 1.18 as follows

Table 1.18

$(a_3 + a_4) / 2$	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8
c_1	0	3	0	3	3	3	1/2	5/2
c_2	3	0	3	1/2	1/2	0	5/2	1/2
c_3	0	3	0	3	3	3	1/2	5/2
c_4	3	1/2	3	0	0	1/2	3	1/2
c_5	3	1/2	3	0	0	1/2	3	1/2
c_6	3	0	3	1/2	1/2	0	5/2	1/2
c_7	1/2	5/2	1/2	3	3	5/2	0	3
c_8	5/2	1/2	5/2	1/2	1/2	1/2	3	0

Using Table 1.10, 1.11 and 1.12 to find Table 1.19 as follows

Table 1.19

$a_1+a_2+a_3/3$	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8
c_1	0	6.4	4.73	7.33	1.73	7.57	0.23	1.67
c_2	6.4	0	1.67	1.07	4.67	1.17	6.5	4.73
c_3	4.73	1.67	0	2.73	6.33	2.83	4.83	6.4
c_4	7.33	1.07	2.73	0	5.73	0.23	7.57	5.67
c_5	1.73	4.67	6.33	5.73	0	5.83	1.83	0.07
c_6	7.57	1.23	2.9	0.23	5.9	0	7.33	5.9
c_7	0.23	6.57	4.9	7.57	1.9	7.33	0	1.9
c_8	1.67	4.67	6.33	5.67	0	5.9	1.9	0

4. Conclusion

This Paper discussed two of the approaches for the determination of the similarity between records of mixed data. We introduce in this paper some concept's and application, from the introduce application, we found that the relation between the general topology and the rough set. From the last, we heard that topology is father of rough set, but in this thesis

showed this relation. In the next paper we will this relation becomes fact. It may be noticed that because of the uncertainty and ambiguity of qualitative data and of trials to combine metrics leave rough set theory as an optional tool to be used. As mentioned in the discussion, an extra or another approach is required for the discovery of identical sets of records in data sets of mixed data.

From Tables, we can notice that the cluster that includes the attributes, the most possible record in the same cluster would be attributing as it is in both approximations. One might use the union of the upper approximations to determine probable clusters.

References

Han, J., & Kamber, M. (2001). *Data Mining concepts and Techniques*, Morgan Kaufmann, San Francisco.

Lin, T. Y., Yao, Y. Y. & Zadeh, L. A. (Eds.) (2002). Rough Sets. *GranularComputing and Data Mining*, Physica-verlag, Heidelberg.

Pawlak, Z. (1982). Rough sets. *International Journal of Computer and Information Science*, 341-356.

Pawlak, Z. (2002). Rough sets and intelligent data analysis. *Information Sciences*, 1-12.

Yao, Y. Y. (1999). Granular computing using neighborhood systems, in: Roy, R., Furuhashi, T . and Chawdhry, P.k. (Eds.), Advances in Soft Computing: Engineering Design and Manufacturing, Springer-Verlag, New York, 539-553.

Zhu, Y. (2002). Unsupervised Database Discovery Based on Artificial Intelligence Techniques. Master's Thesis, University of Cincinnati, June, 1-107.

Copyrights