BEHAVIOUS BASED COI

SING ROUGH SET

WANG LE WEI

A proposal submitted in partial fulfilment of the requirements for the award of the degree of Bachelor of Computer Science (Software Engineering)

FACULTY OF COMPUTER SYSTEM & SOFTWARE ENGINEERING

UNIVERSITI MALAYSIA PAHANG

May, 2013

ABSTRACT

In the context of information technology nowadays, there are many data were emerged when people are using computers, we called it computer user behavior. All of this data are scrambled over inside the computer such as user behavior log files. The problem with this is, when we want to know the user behaviors on computer and doing analysis for specific proposes, we normally needed the data only such as program name and opening time, there are too many to look for and they are all scrambled in log files. Therefore, there are techniques that are proposed that will provide a way to automatically mine the data and obtain only meaningful data from the huge data over the internet. The area discussed in this research is Knowledge Discovery in Databases (KDD) and the technique used is Minimum-Minimum Roughness (MMR). The dataset used will be the dataset of computer user log files. By using this MMR technique, I intended to cluster the user log files dataset which each cluster will contain the data most related to each other.

Contents

CHAPTER 1

INTRODUCTION

This chapter would first talk about the overview of data mining research. There are six parts contained in. The background is the first; then the problem statement. Followed the motivation, next the scopes; the objectives of this research topic would be the last part the thesis organization which briefly describes.

1.1 Background

The concept of KDD is Knowledge Discovery from Databases. That is a kind of program can extract of implicit, previously unknown, and potentially useful information from data. KDD actually provide some kinds of automated methods to make the useful information is mined in databases.

Then refer to Data Mining, Data Mining is a step in the KDD process. It consists both of applying data analysis and discovery algorithm. Also it will produce a calculation method of patterns around the data. And the Data mining methods two of them are Classification and clustering.

A very promising area for attaining this objective is the use of data mining Based on the Romero & Ventura (2006) said that most of the researchers have begun to investigate various data mining methods to help lecturers to improve the learning systems in the last few years. As we know, the data mining also known as a knowledge discovery in databases KDD) is that used by to explore the unique type of data that is unique from the educational context with automatic illustrate and Attractive patterns from big data collections. Data mining is a multidisciplinary area in which several computer paradigms converge and some of the most useful data mining tasks and methods are statist, visualization, clustering, classification and

association rule mining. Data mining can record any student activities such as taking tests and performing various tasks that stores all the information's in the database.

1.2 Problem Statement and Motivation

Computer nowadays becomes comment in people's life and it can be used doing various kinds of tasks. User behaviours log file is a type of text file that recorded details using report from computer user. From opening computer to shutting it down, log file would record every event of computer did. So the log file is usually in a big size as recording huge log information. Now what if I want to find or filter a specify software that user commonly used?

It is very difficult to manage to the huge and big quantities of data manually. We need tools to help us to handle and extract useful information when there are a great number of data log, so the best suggestion is the use of data mining for attaining this objective. To filter and cluster those dataset, I will use data mining roughness theory to assist in the clustering of log file classified base on certain criteria.

Data mining tools are normally designed more for power arid flexibility than for simplicity. My analysis of user log files processed based on filtering program name and opening time, finding implication information such as how frequently people open any software and for how long, as well as the types and sequences of activities that users conduct on these softwares.

By using data mining for database courses can provide the better decision making information. In summary, my analysis demonstrates the power of using Min-Min rough set theory in clustering computer user behaviours.

1.3 Objectives

The objective of this research is to make improvement of data analysis skills and prediction an evaluating results base the data mining results for understanding how users behave when they open computer and launch software or application for better understanding user behaviour patterns. The objectives of this study are:

a. To divide the users and software into some classes base on the rough set rules and theories.

b. To develop a software to implement computer user clustering with the programming language Visual Basic.

c. To validate the method on a real computer user behavior log data set.

1.4 Scopes

There are following scopes of this study are:

a. For dataset using, there is a random collective data which I download from internet include 1000 computer users and their behavior log file during four weeks : 2012-05-07 to 2012-08-13.

b. The clustering uses rough set and Min-Min roughness clustering technique.

c. The minimum roughness and splitting point program will be decided based on the result of user clustering from behaviour log file.

1.5 Thesis Organization

The following chapter will be organized as: Chapter II is literature reviews which describe the Data mining and Educational Data mining. Chapter III is about the methodnology which describes the theory of rough set and modeling process and min-min roughness data clustering, making an example of calculation. Describing the dataset, Chapter IV describes the expected results from an application following by discussion. Finally, Chapter V is the conclusion of this paper.

CHAPTER 2

LITERATURE REVIEW

This chapter will talk about the topic base on some existing literature related with the topic. This chapter contain three sections. The first discuss Knowledge Discovery in Databases. The second describes the concept of Data Mining. The third describes Educational Data Mining.

2.1.1 Definitions of KDD

Kiosgen, W., &Zytkow, J. (2002) defined that knowledge discovery in databases (KDD) is the automatic pulling out of implied and attractive patterns from large data collections.

However, F. HrudayaKu.Tripathyct.al (2007) mentioned the Knowledge Discovery from Databases (KDD) is regularly a multiphase process involving various steps, similar to data preparation, preprocessing, search for hypothesis generation, model of formation, knowledge appraisal, demonstration, alteration and administration. Carter ct al. proposed that knowledge discovery from databases (KDD) is the non-trivial extraction of implicit, previously unknown, and potentially useful information from data. As a branch of machine learning, KDD encompasses a number of automated methods whereby useful information is mined from data stored in databases. When a KDD method is implemented as a practical tool for knowledge discovery in databases, an important requirement is that it be as efficient as possible so that it can handle the large input data sets typically encountered in commercial

environments. This paper I want to present the results of comparing implementations of three similar KDD algorithms to determine their suitability for application to large scale commercial databases.

The technology of build modern database are quickly rising volume and summarize the information they contain increasingly. Knowledge Discovery in Databases (KDD) and data mining are one of the new study areas that try to slove the real world problems.

The objective of doing KDD is to build the pattern of data and show the results which are understandable to humans. The discovered model need to be valid, novel and useful on database result. Using KDD the application would exciting more discovery and higher level information from the datasets in databases even in different territories.

2.1.2. KDD Processes

The definition of KDD process is using the data compare other items, processing, applying data mining methods to enumerate patterns from it. The whole KDD process contain evaluating and determine the interpretations which patterns can be regard as new knowledge.

The process or KDD consists of following steps:

a. Developing an understanding dataset domain and objectives of the client.

b. Selecting a dataset be minged: selecting a dataset which would be processed and shown the results.

c. Data decrease and projection: depending the goal of task, rinsing the useful features to represent the data. To find the invariant representations of the data, the dimensionality decrease or alteration to reduce efficient number would be reflecting.

d. Matching the objective patterns to data mining method such as categorization, decoration or clustering. Choosing the data mining algorithms and techniques for searching the data patterns.

e. Investigative analysis and model and hypothesis selection: choosing the data mining algorithms and selecting methods for searching the data patterns. It includes deciding which models and parameters would be appropriate and matching in the data mining technique with the overall KDD process.

f. Data mining: look for patterns form a set of such representations, including categorization rules clustering. The user would aid the data mining method by performing the preceding steps correctly.

g. Measurement on the knowledge found: using direct knowledge and combine knowledge into other system for further action, or just making report to the parties concerned. This process involves inspection and resolve conflictions will occur with the knowledge that is extracted before and trusted.

2.2.1. Definitions of DM

Srivastava, Cooley, Deshpande, & Tan, (2000) proposed that data mining is a step in the overall process of KDD that consists of preprocessing, data mining and post processing. Data mining has already been successfully applied in e-commerce. However, KDD is the whole process of discovering information from dataset, and KDD is one steps of data mining. In KDD process data mining is the most important part. Data mining uses the particular algorithms to looking for the hidden relation amount the pattern from huge dataset stored in database.

Generally, a algorithm of data mining involves following parts that always combined:

A. The Pattern: it may contain parameters or variable that are to be evaluate from dataset.

B. The Preference criterion: The criterion usually come form of goodness-of-fit function of the model to the data, maybe tempered by a generating a model with many degrees of freedom can be constrained by the given data.

C. The Search Algorithm: the specification of an algorithm for finding the goal models and parameters, giving the data, models and preference criterion.

A particular data mining algorithm is a components. The more familiar model functions in include the following an instantiation of the model existing data mining application

a. Classification: classifies a data item into one of several predefined categorical classes.

b. Regression: maps a data item to actual valued forecast variable.

c. Clustering: maps a data item into one of several clusters, where clusters are natural grouping of data items based in similarity metrics or probability solidity models.

d. Discovering association rules: describes association relationship among different attributes.

e. Summarization: provides a compact description for a subset of data.

2.3. User Behaviour Data Mining

User behaviour analysis and data mining can be applied from different types of territories such as user behaviours in internet; in wireless LAN network or in computer system. It is necessary to deal separately with the application of- data mining techniques in each type due to the fact that they have different data sources and objectives.

2.3.1. User Behavior in Internet (example as Online Social Networks)

Online social networks (OSNs) have become great popular in humans life. According to Nielsen Online's latest research（2009）, social media have pulled ahead of email as the most popular online activity. More than two-thirds of the global online population visit and participate in social networks and blogs. In fact, social networking and blogging account for nearly 10% of all time spent on the Internet.

These statistics suggest that OSNs have become a fundamental part of the global online experience. Through OSNs, users connect with each other, share and content, and disseminate information. Numerous sites provide social links, for example, networks of professionals and contacts (e.g., LinkedIn, Facebook, MySpace) and networks for sharing content (e.g., Flickr, YouTube).

Motivation of studying how users behave when they connect to these sites:

First, studies of user behaviors allow the performance of existing systems to be evaluated and lead to better site design (M. Burke, C. Marlow, and T. Lento. 2009) and advertisement placement policies(B. A. Williamson. 2007).

Second, accurate models of user behavior in OSNs are crucial in social studies as well as in viral marketing. For instance, viral marketers might want to exploit models of user interaction to spread their content or promotions quickly and widely (J. Leskovec, L. A. Adamic, and B. A. Huberman. 2007).

Third, understanding how the workload of social networks is re-shaping the Internet traffic is valuable in designing the next-generation Internet infrastructure and content distribution systems (B. Krishnamurthy. 2009). Despite the potential benefits, little is known about social network workloads. A few recent studies examined the patterns using data that can be gathered from OSN sites, for instance, writing messages to other users (B. Huberman, D. Romero, and F. Wu. 2009) or accessing third party applications (A. Nazir, S. Raza, and C.-N. Chuah. 2008).

Method of study user behaviours on OSN workloads:

A complementary approach to study OSN workloads is to use traces such as clickstream data that capture all activities of users (P. Chatterjee, D. L. Ho_man, and T. P. Novak. 2003). Since clickstream data include not only visible interactions, but also silent" user actions like browsing a page or viewing a photo, they can provide a more accurate and comprehensive view of the OSN workload.

There is a kind analysis of OSN workloads based on a clickstream dataset collected from a social network aggregator. Social network aggregators are one-stop shopping sites for OSNs and provide users with a common interface for accessing

multiple social networks(R. King. 2007).

By using the clickstream data, there should be conducted three sets of analyses:

- First, characterized the traffic and session patterns of OSN workloads and examined how frequently people connect to OSN sites and for how long. Based on the data, providing a best fit models of session inter-arrival times and session length distributions.

- Second, using the analysis strategy called the clickstream model, to characterize user activity in OSNs. The clickstream model captures dominant user activities and the transition rates between activities.

In summary, the clickstream data analyzed in the paper provides an accurate view of how users behave when they connect to OSN sites. Furthermore, the data analysis could suggest several interesting insights into how users interact with friends in OSNs. And the findings of study result could have implications for efficient system design.

2.3.2. User Behavior in Wireless LAN Network

Wireless LAN installations based on IEEE 802.11 (IEEE. 802.11b/d3.0 Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specification, 2009) technology are emerging as an attractive solution for providing network connectivity in corporations and universities, and in public places like conference venues, airports, shopping malls, etc. – places where individuals spend a considerable amount of their time outside of home and work. In addition to the convenience of untethered networking, contemporary wireless LANs provide relatively high data connectivity at 11 Mb/s and are easy to deploy in public settings.

Motivation of studying

The high-level goals of this study are two-fold:

First, specifically characterize user behavior and network performance in a public wireless LAN environment. Characterize user behavior in terms of connection session length, user distribution across APs, mobility, application mix, and bandwidth requirements and it is also necessary to characterize network performance in terms of overall and individual AP load, and packet errors and retransmissions. From these analyses wireless users in terms of a parameterized

model for use with analytic and simulation studies involving wireless LAN traffic can be characterize.

Second, to better understand the issues in wireless network deployment, such as capacity planning, and potential network optimizations, or algorithms for load balancing across multiple APs in a wireless network.

Method of study user behaviors on OSN workloads:

In this study, it support to use a trace recorded method to present and analysis user behavior and network performance in a public-area wireless network. The trace consists of two parts.

The first part is a record of performance monitoring data sampled from wireless access points (APs) serving the conference,

The second consists of anonymized packet headers of all wireless traffic.

CHAPTER 3

Methodology

This chapter discusses about the model and method of data clustering based on Rough Set Theory. It include information system and rough set theory; Set approximations and indiscernibility relations; Min-min roughness technique with an example of the student evaluation dataset. Then as last shows the splitting model.

3.1 Rough Set Theory

Rough set theory has attracted many researchers all over the world who contributed essentially to its develop the applications. The objective of the rough set theory is induction of approximations of concepts. The idea consists of approximation of a subset by a pair of two precise concepts called the lower approximation and upper approximation. Intuitively, the lower approximation of a

set consists of all elements that surely belong to the set, whereas the upper approximation of the set constitutes of all elements that possibly belong to the set. The difference of the upper approximation and the lower approximation is a boundary region. It consists of all elements that cannot be classified uniquely to the set or its complement, by employing available knowledge. Thus any rough set, in contrast to a crisp set, has a non-empty boundary region.

3.1.1 Information System

Data are often presented as a table, columns of which are labeled by attributes, rows by objects of interest and entries of the table are attribute values. By an information system, an information system is also called a knowledge representation systems or an attribute-valued system and can be intuitively expressed in terms of an information table (refer to Table 3.1).

U	a_1	a_2	$\cdots$	a_k	$\cdots$	$a_{\lvert A\rvert}$
u_1	$f(u_1, a_1)$	$f(u_1, a_2)$	$\cdots$	$f(u_1, a_k)$	$\cdots$	$f(u_1, a_{\lvert A\rvert})$
u_2	$f(u_2, a_1)$	$f(u_2, a_2)$	$\cdots$	$f(u_2, a_k)$	$\cdots$	$f(u_2, a_{\lvert A\rvert})$
u_3	$f(u_3, a_1)$	$f(u_3, a_2)$	$\cdots$	$f(u_3, a_k)$	$\cdots$	$f(u_3, a_{\lvert A\rvert})$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
$u_{\lvert U\rvert}$	$f(u_{\lvert U\rvert}, a_1)$	$f(u_{\lvert U\rvert}, a_2)$	$\cdots$	$f(u_{\lvert U\rvert}, a_k)$	$\cdots$	$f(u_{\lvert U\rvert}, a_{\lvert A\rvert})$

Table 3.1: An information system

Example 3.1.

There is a **student final result list from Moodle dataset,** as shown in Table 3.2.

Student/ coursework	**Internet Explorer (A1)**	**MSN (A2)**	**Microsoft Office (A3)**	**Photoshop (A4)**	**Adobe Reader (A5)**	**Game (A6)**	**Grade**
Log file1 (U1)	8	7	6	8	9	7	A
Log file2 (U2)	9	7	7	8	8	8	A
Log file3 (U3)	7	5	7	8	6	5	C
Log file4 (U4)	9	6	5	6	5	7	B
Log file5 (U5)	6	7	5	7	6	6	B
Log file6 (U6)	7	7	6	7	6	6	C

Table 3.2: A **student evaluation system**

For easy writing and understanding. The mark regard as the students matric No and assessments should be cleared. The values for calculation in following will be gained from Table 3.2.

U = {U1, U2, U3, U4, U5, U6}

A = {A1, A2, A3, A4, A5, A6, Grade}, where C = {A1, A2, A3, A4, A5, A6}, D = {Grade},

VA1 = {6,7,8,9}

VA2 = {6,7,8}

VA3 = {5,6,7}

VA4 = {6,7,8}

VA5 = {5,6,8,9}

VA6 = {5,6,7,8}

VGrade = {A, B, C}

3.1.2 Set Approximations

Definition 2.2. *Let* $S = (U, A, V, f)$ *be an information system, let B be any subset of A and let X be any subset of U. The B-lower approximation of X, denoted by* $\underline{B}(X)$ *and B-upper approximations of X, denoted by* $\overline{B}(X)$, *respectively, are defined by*

$$\underline{B}(X) = \{x \in U \mid [x]_B \subseteq X\} \text{ and } \overline{B}(X) = \{x \in U \mid [x]_B \cap X \neq \phi\}.$$

a. The *lower approximation* of a set X Compared with B is the set of all objects, which can be for *certain* classified as X using B (are certainly X in view of B).

b. The *upper approximation* of a set X Compared with B is the set of all objects which can be *possibly* classified as X using B (are possibly X in view of B).

The accuracy of approximation (accuracy of roughness) of any subset $X \subseteq U$ Compared with $B \subseteq A$, denoted $\alpha_B(X)$ is measured by

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

where $|X|$ denotes the cardinality of X. For empty set ϕ, it is defined that $\alpha_B(\phi) = 1$ (Pawlak and Skowron, 2007). Obviously, $0 \leq \alpha_B(X) \leq 1$. If X is a union

of some equivalence classes of U, then $\alpha_B(X)=1$. Thus, the set X is crisp (precise) Compared with B. And, if X is not a union of some equivalence classes of U, then $\alpha_B(X)<1$. Thus, the set X is rough Compared with B (Pawlak and Skowron, 2007).

3.2 Min-Min Roughness

There are a few techniques that had been proposed to deal with the clustering attribute selection. Mazlack et al. proposed two techniques to select clustering attribute, which is bi-clustering (BC) technique and total roughness (TR) technique. Then, Parmaret al. proposes a technique called min-min roughness (MMR) which improves the BC technique for data set with multi-valued attributes. Another techniques also had been proposed, which is called maximum dependency of attributes (MDA).

3.2.1 Model for selecting a clustering attribute

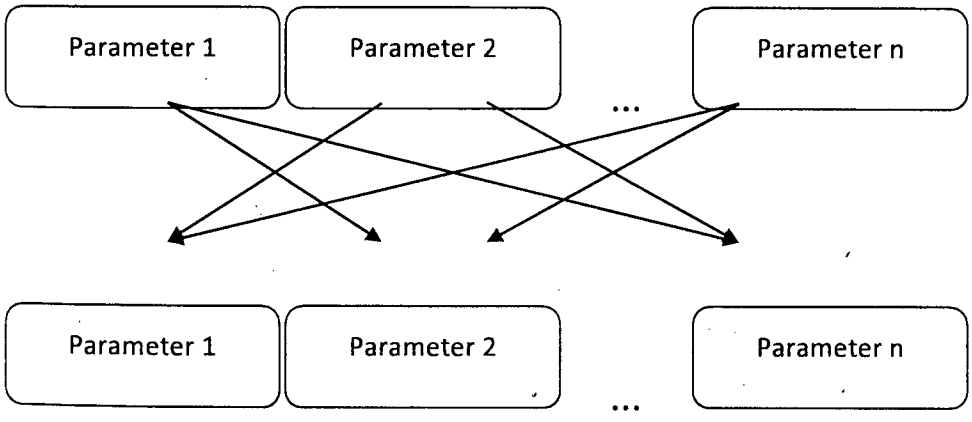

Figure 3.2: Selecting a clustering model.

3.2.2 Min-Min Roughness Technique

The following Table shows step-by-step to calculate Min-Min Roughness.

Step	Min-Min Roughness
1	Given data set
2	Each attribute in data set considered as a candidate attribute to partition
3	Determine equivalence classes of attribute-value pairs
4	Determine lower approximation of each equivalence classes in attribute a_i w.r.t. to attribute a_j, $i \neq j$
5	Determine upper approximation of each equivalence classes in attribute a_i w.r.t. to attribute a_j, $i \neq j$
6	Calculate roughness of each equivalence classes in attribute a_i w.r.t. to attribute a_j, $i \neq j$
7	Calculate mean roughness of attribute a_i w.r.t. to attribute a_j, $i \neq j$
8	Calculate minimum roughness a_i w.r.t. to all attribute a_j, $i \neq j$
9	If there are two greatest value of mean roughness, calculate minimum roughness relative to the second, third greater minimum roughness until the tie is broken
10	Selecting a clustering attribute

Table 3.3: Step-by-step Min-Min Roughness

3.2.3 Algorithm

Below show the example algorithm to obtain the MMR from the students information dataset.

// finding the *U/IND* for each attribute

x=1
for i = 1 to nth attribute
    set att(i, 1) as set(i,x)
    for j = 1 to nth row
        for k = 1 to nth
        if att(i,j) doesn't belong to any set & $j \neq k$
            thenif att(i,j) = att(i,k) & att(i,k) belong to a set
                then set att(i,j) as same set as att(i,k)
                else set att(i,j) as set(i,x++)
            end if
        end if
    end for loop
end for loop

// finding the number of element in lower and upper approximation for each attribute

for i = 1 to nth attribute
    for j = 1 to nth attribute
        for k = 1 to nth attributeSet
            if set(j,k) $\in$ set(i,k) & $i \neq j$
                then lowerApprox(a_i,k) =
                    lowerApprox(a_i,k) + no of element in
                    set(j,k)
            else if set(i,k) $\in$ & ij
                then upperApprox(a_i,k) =
                    upperApprox(a_i,k) + no of element in
                    set(j,k)
            end if
        end for loop
    end for loop
end for loop

// calculating roughness for each attribute set

for i = 1 to nth attribute
    for k = 1 to nth attributeSet
        roughness(a_i,k) =
        1 – (lowerApprox(a_i,k) ÷ upperApprox(a_i,k))
    end for loop
end for loop

// calculating mean roughness for each attribute

x=1
for i = 1 to nth attribute
    for k = 1 to nth attributeSet
        totalRoughness() =
        totalRoughness(a_i) + roughness(a_i,k)
        totalAttritbute(a_i) = totalAttritbute(a_i) + x
    end for loop
end for loop

for i = 1 to nth attribute
    meanRoughness(a_i) = totalRoughness(a_i) ÷ totalAttritbute(a_i)
end for loop

// finding the min roughness from all attributes

for i = 2 to nth attribute
    if meanRoughness(a_i) < meanRoughness(a_{i-1})
        then minRoughness = meanRoughness(a_i)
    end if
end for loop

// for case where the lowest minimum meanRoughness is more than 1

count = 0
for i = 1 to nth attribute
    if minRoughness = meanRoughness(a_i)
        then count = count + 1
    end if

```
end for loop
if count > 1
    thenfor  i = 2 to nth attribute
             if   meanRoughness(a_i) < meanRoughness(a_i-1)
                  & meanRoughness(a_i)  ≠  minRoughness
                  then     min-minRoughness =
                       meanRoughness(a_i)
             end if
        end for loop
    count = 0
    for i = 1 to nth attribute
        if   min-minRoughness = meanRoughness(a_i)
             then     count = count + 1
        end if
    end for loop
end if
```

Calculate approximations and roughness

First, we determine of indiscernibility relation of singleton attribute are:

S (A1 = 6) = {(U5)}, S(A1 = 7) = {U3, U6} S(A1 = 8) = {U1} S(A1 = 9) = {U2, U4}

S(A2 = 5) = {U3}, S(A2 = 6) = {U4}, S(A2 = 7) = {U1, U2, U5, U6}

S(A3 = 5) = {U4, U5}, S(A3 = 6) = {U6}, S(A3 = 7) = {U2, U3}

S(A4 = 6) = {U4}, S(A4 = 7) = {U5, U6}, S(A4 = 8) = {U1, U2, U3}

S(A5 = 5) = {U4}, S(A5 = 6) = {U3, U5, U6}, S(A5 = 8) = {U2}, S(A5 = 9) = {U1}

S(A6 = 5) = {U3}, S(A6 = 6) = {U5, U6}, S(A6 = 7) = {U1, U4}, S(A6 = 8) = {U2}

Then we find the upper and lower approximations.

Attribute A1

For attribute A1, shown that |V(A1)|=4). The approximations and roughness about A_1 Compared A_i which i= 2,3,4,5,6, calculated as the following.

1) Compared with A2

lower app and upper app calculated:

$\underline{X(A1=6)} = \varnothing$ and $\overline{X(A1=6)} = \{U1 \neg U2 \neg U5 \neg U6\}$

$\underline{X(A1=7)} = \{U3\}$ and $\overline{X(A1=7)} = \{U1 \neg U2 \neg U3 \neg U5 \neg U6\}$

$\underline{X(A1=8)} = \varnothing$ and $\overline{X(A1=8)} = \{U1 \neg U2 \neg U5 \neg U6\}$

$\underline{X(A1=9)} = \{U4\}$ and $\overline{X(A1=9)} = \{U4\}$

Roughness

$RA2(S \mid A1=6) = 1-0/4 = 1$
$RA2(S \mid A1=7) = 1-1/5 = 0.8$
$RA2(S \mid A1=8) = 1-0/4 = 1$
$RA2(S \mid A1=9) = 1-1 = 0$

Mean roughness

$$\text{Rough A2(A1)} = (1+0.8+1+0)/4 = 0.7$$

2) Compared with a_3

lower app and upper app calculated: