การใช้ทฤษฎี Rough Set เพื่อการวิเคราะห์ข้อมูลอัตโนมัติ*

Using Rough Set Theory for Automatic Data Analysis

วราภรณ์ โปตะวัฒน์, คมศัลล์ ศรีวิสุทธิ์, อุษารัตน์ แสนปากดี, สุนิดา ศรีสุริยชัย, นิตยา เกิดประสพ, กิตติศักดิ์ เกิดประสพ

Varaporn Potawat, Komsan Sriwisut, Usarat Sanpakdee, Sunida Srisuriyachai, Nittaya Kerdprasop, Kittisak Kerdprasop

School of Computer Engineering, Suranaree University of Technology, 111 University Ave., Muang District, Nakhon Ratchasima 30000, Thailand. E-mail address: kerdpras@ccs.sut.ac.th

บทคัดย่อ: ระบบสารสนเทศเป็นแหล่งรวมข้อมูล เหตุการณ์ และกรณีศึกษาต่างๆ ที่สามารถนำมาค้นหาความรู้ที่เป็นประโยชน์ได้ ความรู้ที่ได้จะอยู่ในลักษณะของการสัมพันธ์เชื่อมโยงระหว่างปัจจัยที่เป็นต้นเหตุของเหตุการณ์หนึ่ง และผลที่ทำให้เกิดอีกเหตุการณ์หนึ่งขึ้น งานวิจัยนี้เป็นการทดลองนำทฤษฎี Rough Set มาใช้ในการวิเคราะห์ระบบสารสนเทศของนักศึกษาเพื่อค้นหากฎที่สามารถนำมาใช้ช่วยคาดหมายสถานภาพของนักศึกษาโดยพิจารณาจากประวัติและผลการเรียนในรายวิชาต่างๆ

Abstract: The information system that contains records about particular cases, events, objects, or observations is a valuable source of knowledge. Each record is consisted of two kinds of attributes: condition and decision attributes. Condition attributes concern results of some tests or measurements, data from observations, symptoms of cases, etc. Decision attributes concern some expert's decisions, diagnoses, classified results of a treatment, etc. The question about cause-effect dependencies between condition-decision attributes is the most interesting issue in the analysis of information systems. We employ the rough set approach as an analysis tool on the student information. The main purpose of our experimentation is to investigate the efficiency of rough set theory to the real-world data analysis problem. The result obtained from the rough set analysis is a set of decision rules describing the dependencies between some conditions to the status of the students.

Introduction: In the recent years we observe growing interest in rough set theory and its applications all over the world. Rough set theory was introduced by Zdzislaw Pawlak in the early 80's [5,6]. The main issue of rough set theory is reasoning from imprecise data. The concept of rough set comes from the notion of indiscernibility (or similarity) of knowledge. An indiscermibility relation is an equivalence relation that identifies objects which have the same descriptions with respect to a set of attributes of objects. The indiscernibility relation can be employed in order to define approximations of sets or relations. The notion of indiscernibility relation is useful for expressing several interesting properties of an incomplete information system. The rough set method allows us to discover the knowledge in the form of optimal decision rules. The method is advantageous for a very large data set and has been successfully applied to a lot of applications in various fields such as medicine, finance, telecommunication, process industry, marketing, etc [2,4,7,8].

ROSE [1], RSES and ROSETTA [3] are some examples of rough set software developed for solving knowledge discovery problems. ROSE is a modular software system implementing basic elements of the rough set theory and rule discovery techniques. It has been created at the Laboratory of Intelligent Decision Support Systems in Poland. It has many tools to use for Rough set based knowledge discovery such as analyzing data, extracting characteristics patterns from data, including decision rules from sets of learning examples and evaluating the discovered rules.

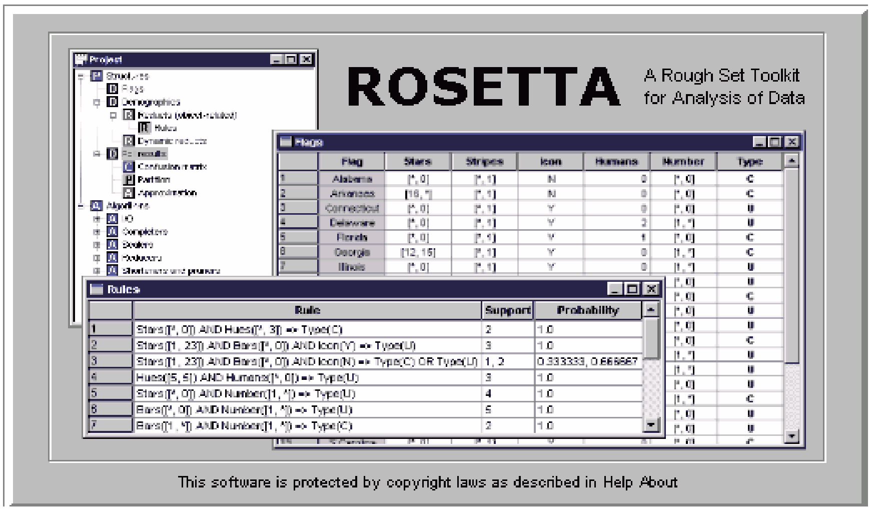
RSES system was developed at Warsaw University. A subsequent system named ROSETTA has been designed and implemented at the Norwegian University of Science and Technology in Trondheim, Norway. The computational kernel of ROSETTA is a C++ class library for performing rough set related calculation to discover knowledge. The discovery results are propositional rules synthesized from empirical data.

Methodology: The aim of this research is to investigate the power of rough set theory in discovering models or patterns that describe the real-world data. The data set used in our experiments is the student data of Suranaree University of Technology (SUT) during the year 2001-2002. The two rough set systems chosen as an analysis tool are ROSE2 (Lite version) and ROSETTA. The major steps in data analysis are composed of transforming data to the specified format, fill in the missing values with mean/mode replacement technique, discretization,

* ได้รับทุนสนับสนุนการวิจัยจากศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ รหัสโครงการ NT-B-06-4C-19-524

pattern induction. The results are patterns expressed as rules to describe the norm characteristics of the SUT students. The discovered rules based on rough set induction are cross-validated with he rules obtained from decision tree induction, which is the major technique normally used in data mining.

Results : Figures 1 and 2 show the screenshots of the software ROSETTA and ROSE, respectively. The (part of) results obtained as a set of reducts are shown in Figure 3.

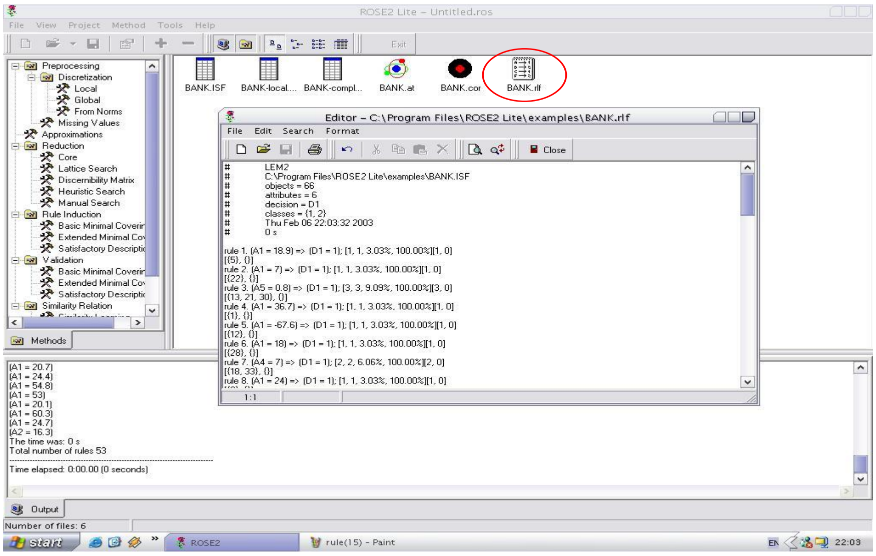


	Flag	Stars	Stripes	Icon	Humans	Number	Type
1	Alabama	[*, 0]	[*, 1]	N	0	[*, 0]	C
2	Arkansas	[16, *]	[*, 1]	N	0	[*, 0]	C
3	Connecticut	[*, 0]	[*, 1]	Y	0	[*, 0]	U
4	Delaware	[*, 0]	[*, 1]	Y	2	[1, *]	U
5	Florida	[*, 0]	[*, 1]	Y	1	[*, 0]	C
6	Georgia	[12, 16]	[*, 1]	Y	0	[1, *]	C
7	Illinois	[*, 0]	[*, 1]	Y	0	[1, *]	U

	Rule	Support	Probability
1	Stars([*, 0]) AND Hues([*, 3]) => Type(C)	2	1.0
2	Stars([1, 23]) AND Bars([*, 0]) AND Icon(Y) => Type(U)	3	1.0
3	Stars([1, 23]) AND Bars([*, 0]) AND Icon(N) => Type(C) OR Type(U)	1, 2	0.333333, 0.666667
4	Hues([5, 5]) AND Humans([*, 0]) => Type(U)	3	1.0
5	Stars([*, 0]) AND Number([1, *]) => Type(U)	4	1.0
6	Bars([*, 0]) AND Number([1, *]) => Type(U)	5	1.0
7	Bars([1, *]) AND Number([1, *]) => Type(C)	2	1.0

This software is protected by copyright laws as described in Help About

Fig. 1. The ROSETTA system



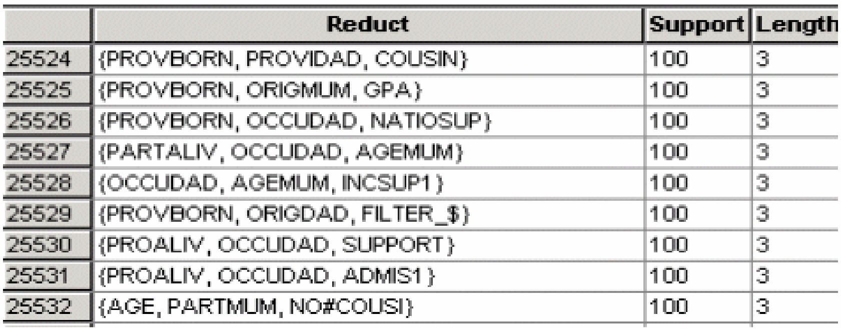
```
#          LEM2
#          C:\Program Files\ROSE2 Lite\examples\BANK.ISF
#          objects = 66
#          attributes = 6
#          decision = D1
#          classes = {1, 2}
#          Thu Feb 06 22:03:32 2003
#          0 s

rule 1. (A1 = 18.9) => (D1 = 1); [1, 1, 3.03%, 100.00%][1, 0]
[{5}, {}]
rule 2. (A1 = 7) => (D1 = 1); [1, 1, 3.03%, 100.00%][1, 0]
[{22}, {}]
rule 3. (A5 = 0.8) => (D1 = 1); [3, 3, 9.09%, 100.00%][3, 0]
[{13, 21, 30}, {}]
rule 4. (A1 = 36.7) => (D1 = 1); [1, 1, 3.03%, 100.00%][1, 0]
[{1}, {}]
rule 5. (A1 = -67.6) => (D1 = 1); [1, 1, 3.03%, 100.00%][1, 0]
[{12}, {}]
rule 6. (A1 = 18) => (D1 = 1); [1, 1, 3.03%, 100.00%][1, 0]
[{28}, {}]
rule 7. (A4 = 7) => (D1 = 1); [2, 2, 6.06%, 100.00%][2, 0]
[{18, 33}, {}]
rule 8. (A1 = 24) => (D1 = 1); [1, 1, 3.03%, 100.00%][1, 0]
```

```
(A1 = 20.7)
(A1 = 24.4)
(A1 = 54.8)
(A1 = 53)
(A1 = 20.1)
(A1 = 60.3)
(A1 = 24.7)
(A2 = 16.3)
The time was: 0 s
Total number of rules 53
------------------------------------------------------------
Time elapsed: 0:00.00 (0 seconds)
```

Number of files: 6

Fig. 2. The ROSE system



	Reduct	Support	Length
25524	{PROVBORN, PROVIDAD, COUSIN}	100	3
25525	{PROVBORN, ORIGMUM, GPA}	100	3
25526	{PROVBORN, OCCUDAD, NATIOSUP}	100	3
25527	{PARTALIV, OCCUDAD, AGEMUM}	100	3
25528	{OCCUDAD, AGEMUM, INCSUP1}	100	3
25529	{PROVBORN, ORIGDAD, FILTER_$}	100	3
25530	{PROALIV, OCCUDAD, SUPPORT}	100	3
25531	{PROALIV, OCCUDAD, ADMIS1}	100	3
25532	{AGE, PARTMUM, NO#COUSI}	100	3

Fig. 3. The results expressed as a set of reducts

The first ten common patterns discovered from the student data by the rough set software can be explained in plain English as follows:

(1) IF GPATrimester12 = GE3.0 AND GPATrimester14 = NA AND grade_102101 = C+
THEN Graduated

(2) IF GPATrimester12 = GE3.0 AND GPATrimester14 = NA AND grade_102101 = C
THEN Graduated

(3) IF GPATrimester12 = GE3.0 AND GPATrimester14 = NA AND grade_102101 = B+
THEN Graduated

(4) IF GPATrimester12 = LT2.5 AND Province = Nakhonratchasima
THEN Graduated

(5) IF GPATrimester12 = LT2.5 AND Province = Chaiyapoom
THEN Graduated

(6) IF GPATrimester12 = LT2.5 AND Province = Khonkaen
THEN Graduated

(7) IF GPATrimester12 = LT2.5 AND Province = Bangkok AND grade_103101 = C+
THEN Graduated

(8) IF GPATrimester12 = LT2.5 AND Province = Bangkok AND grade_103101 = C AND GPATrimester14 = LT1.80
THEN Studying

(9) IF GPATrimester12 = LT1.80 AND grade_102103 = F
THEN Quit

(10) IF GPATrimester5 = NA AND grade_402201 = F AND Province = Bangkok
THEN Quit

Conclusion: Rough set theory is a framework for identifying the least decision rules that are discovered from large amount of data. The result of discovering is the knowledge describing interest patterns of data. The significant issues of rough set are discernibility, approximation, reduct and decision rules. We investigate the application of rough set to the real-world data of SUT students and the results turned out to be interesting and contributed to a better understanding of students' behavior. A more thorough investigation and comparison of rough set technique to other discovery techniques are our future research plan.

References:

[1] Laboratory of Intelligent Decision Support Systems (IDSS): ROSE: Rough Sets Data Explorer. Available URL: http://www-idss.cs.put.poznan.pl/software/rose/

[2] Lin, T.Y., Cercone, N.(eds.): Rough sets and data mining. Analysis of Imprecise Data. Kluwer Academic Publishers, Dordrecht (1997)

[3] Ohrn, A.: Disceribility and rough sets in medicine: Tools and applications. PhD Thesis, Norwegian University of Science and Technology, Department of Computer and Information Science (1999). Available URL: http://www.idi.ntnu.no/~aleks/thesis/

[4] Orlowska, E. (ed.): Incomplete information: Rough set analysis. Physical Verlag, Heidelberg (1997)

[5] Pawlak, Z.: Rough sets. International Journal of Information and Computer Science 11 (1982) 344-356

[6] Pawlak, Z.: Rough sets. Theoretical Aspects of Reasoning about Data. Kluwer Academic Publishers, Dordrecht (1991)

[7] Slowinski, R.(ed.): Intelligent Decision Support – Handbook of Applications and Advances of the Rough Sets Theory. Kluwer Academic Publishers, Dordrecht (1992)

[8] Ziarko, W. (ed.): Rough Sets, Fuzzy Sets and Knowledge Discovery (RSKD'93). Workshops in Computing, Springer-Verlag & British Computer Society, London, Berlin (1994)