Empirical Analysis on the Geriatric Care Data Set Using Rough Sets Theory

Jiye Li[1] and Nick Cercone[2]

[1] School of Computer Science, University of Waterloo
200 University Avenue West, Waterloo, Ontario, Canada N2L 3G1
j27li@uwaterloo.ca
[2] Faculty of Computer Science, Dalhousie University
6050 University Avenue, Halifax, Nova Scotia, Canada B3H 1W5
nick@cs.dal.ca

Abstract. A well known problem for association rules generation is that too many rules are generated, and it is difficult to determine manually which rules are more useful, interesting and important. In our study of using rough sets theory to improve the utility of association rules, we propose a new rule importance measure to select the most appropriate rules. In order to explore the application of the proposed method to large data sets, we perform the experiments on a Geriatric Care data set.

1 Motivation

An association rules algorithm helps to find patterns which relate items from transactions. For example, in market basket analysis, by analyzing transaction records from the market, we could use association rules algorithms to discover different shopping behaviors such as, when customers buy bread, they will probably buy milk. Association rules can then be used to express these kinds of behaviors, thus helping to increase the number of items sold in the market by arranging related items properly. A well known problem for association rules generation is that too many rules are generated, and it is difficult to determine manually which rules are more useful, interesting and important. In our study of using rough sets theory (Please refer to Appendix C for related work on rough sets theory.) to improve the utility of association rules, we propose a new rule importance measure to select the most appropriate rules. In order to explore the application of the proposed method to large data sets, we perform the experiments on a geriatric care data set for reduct and core generations.

We use ROSETTA GUI version 1.4.41 rough sets toolkit [1] for multiple reducts generation. The reducts are obtained by Genetic Algorithm and Johnson's Algorithm with the default option of full discernibility [2]. The intersection of all the possible reducts is called the *core*. The core contains the most important information of the original data set. Hu et al.[3] proposed a new core generation algorithm. We use Hu's algorithm to generate core attributes and to exam the effect of core attributes on the generated rules. In this experiment, a geriatric

care data set is used as our test data set. The attributes for this medical data set are listed in Appendix B. We use *survival status* as the decision attribute. Another 44 attributes on the symptoms of a patient about *education level*, *eyesight*, *hearing*, *age* and so on are used as condition attributes. The data set contains 8547 records, and there is no missing value in this data set.

The experiment procedure is shown in Figure 1.

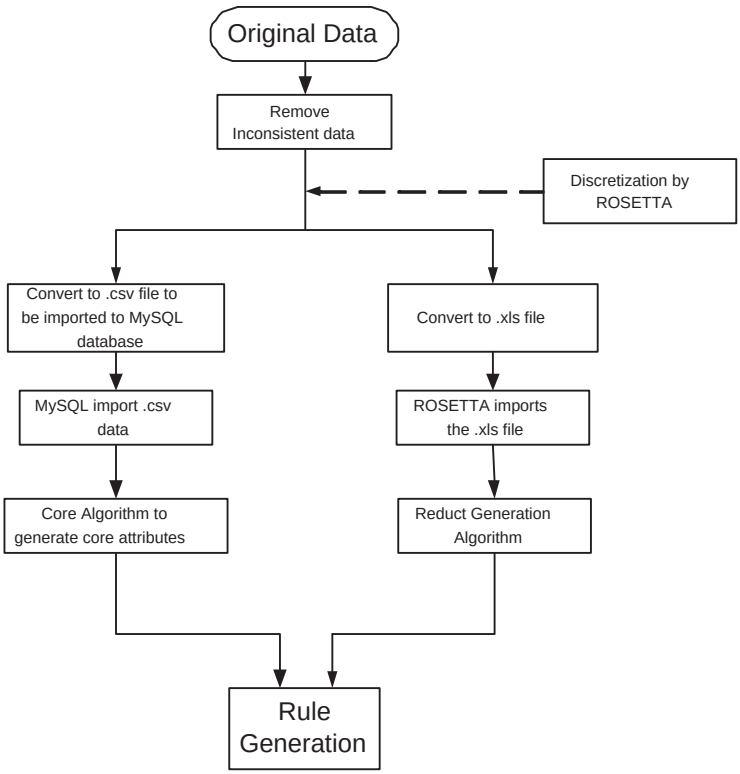

Fig. 1. Experiment Procedure

For the rest of this report, section 2 gives the experiment results on the original data set, without preprocessing. Section 3 contains the core and reduct sets generated after removing inconsistent data from the original data set, as well as the results generated from discretization algorithms. Section 4 shows association rules generated from Johnson's reduct. We summarize our experiments in section 5.

2 Original Data Set

Table 1 gives selected data records of this data set. Our first experiment was

Table 1. Geriatric Care Data Set

edulevel	eyesight	...	health	trouble	livealone	cough	hbp	heart	...	studyage	sex	livedead
0.6364	0.25	...	0.25	0.00	0.00	0.00	0.00	0.00	...	73.00	1.00	0
0.7273	0.50	...	0.25	0.50	0.00	0.00	0.00	0.00	...	70.00	2.00	0
0.9091	0.25	...	0.00	0.00	0.00	0.00	1.00	1.00	...	76.00	1.00	0
0.5455	0.25	...	0.50	0.00	1.00	1.00	0.00	0.00	...	81.00	2.00	0
0.4545	0.25	...	0.25	0.00	1.00	0.00	1.00	0.00	...	86.00	2.00	0
0.2727	0.00	...	0.25	0.50	1.00	0.00	1.00	0.00	...	76.00	2.00	0
0.0000	0.25	...	0.25	0.00	0.00	0.00	0.00	1.00	...	76.00	1.00	0
0.8182	0.00	...	0.00	0.00	0.00	0.00	1.00	0.00	...	76.00	2.00	0
...	...	...	...	...	...	...	...	...	...	...	...	...

Table 2. Core attributes for the original medical data set without preprocessing

Core	Core Attributes
Core	eartroub,meal,chest,bathroom,getbed, studyage,trouble,fracture,shower,nerves, diabetes,walk,dental,age6,dress, eat,sneeze,parkinso,eyetroub,stomach, shopping,takemed,money,edulevel,kidney, walkout,housewk,arthriti,tired,takecare, skin,feet,bowels,eyesight,livealone, hbp,heart,stroke,hearing,sex, health,phoneuse,bladder,cough

to generate the core attributes directly from the original data set without any preprocessing. The core algorithm returns 44 attributes as shown in Table 2. This implies all of the condition attributes are the core attributes.

We use ROSETTA toolkit to generate reducts. The result contains 17 reduct sets as shown in Table 3.

The reduct sets do not contain all the core attributes. We also apply different discretization algorithms provided by ROSETTA toolkit. The result shows inconsistency between the core attributes and the reduct sets.

3 Data Cleaning

The result of our experiment on the original data is not satisfactory. We perform the following preprocessing. We first sort the whole data set according to the condition attributes, excluding the decision attributes. Then we select data entries that contain the same 44 condition attributes, but different decision attributes. This data is inconsistent, because, given the exactly same illness symptoms, the

Table 3. Reduct sets for the original medical data set without preprocessing

Reduct Generation	Reduct Set
Genetic Algorithm No. 1	{edulevel, eyesight, hearing, housewk, health, trouble, livealone, hbp, heart, arthriti, eyetroub, eartroub, dental, chest, stomach, diabetes, feet, nerves, skin, studyage, sex}
No. 2	{edulevel, eyesight, hearing, housewk, health, trouble, livealone, hbp, heart, arthriti, eyetroub, eartroub, dental, chest, kidney, diabetes, feet, nerves, skin, studyage, sex}
No. 3	{edulevel, eyesight, hearing, housewk, health, trouble, livealone, hbp, heart, arthriti, eyetroub, eartroub, dental, chest, stomach, bladder, diabetes, feet, nerves, studyage, sex}
..	...
No. 17	{edulevel, eyesight, hearing, shower, housewk, money, health, livealone, cough, tired, sneeze, hbp, heart, arthriti, eyetroub, eartroub, dental, chest, stomach, bladder, diabetes, feet, nerves, studyage, sex}

Table 4. Core attributes for the medical data set after cleaning

Core	Core Attributes
Core	eartroub,livealone,heart,hbp,eyetroub, hearing,sex,health,edulevel,chest, housewk,diabetes,dental,studyage

decision should be the same. After preprocessing, we found 12 inconsistent data entries in the medical data set. We remove these 12 records. Therefore, the data contains 8535 records.

3.1 Without Discretization

Core attributes are generated based on the data without inconsistent records.

Table 4 shows the 14 core attributes. Table 5 shows the reduct sets generated by ROSETTA.

There are 86 reduct sets generated by genetic algorithm, and all of them contain the core attributes. Johnson's algorithm [3] generates one reduct set and all the core attributes are in the reduct set.

[3] Reduct sets generated by Johnson's algorithm are listed for the purpose of comparison. Because this algorithm only generates one single reduct, we will not use the result in our experiment.

Table 5. Reduct sets for the medical data set after cleaning

Reduct Gen. Algorithm	Reduct Set
Johnson's Algorithm	{edulevel, eyesight, hearing, housewk, health, trouble, livealone, hbp, heart, arthriti, eyetroub, eartroub, dental, chest, bladder, diabetes, feet, nerves, studyage, sex}
Genetic Algorithm No. 1	{edulevel, eyesight, hearing, shopping, housewk, health, trouble, livealone, cough, sneeze, hbp, heart, arthriti, eyetroub, eartroub, dental, chest, kidney, diabetes, feet, nerves, skin, studyage, sex}
No. 2	{edulevel, eyesight, hearing, phoneuse, meal, housewk, health, trouble, livealone, cough, sneeze, hbp, heart, arthriti, eyetroub, eartroub, dental, chest, bladder, diabetes, feet, nerves, skin, studyage, sex}
...	...
No. 86	{edulevel, eyesight, hearing, shopping, meal, housewk, takemed, health, trouble, livealone, cough, tired, sneeze, hbp, heart, stroke, arthriti, eyetroub, eartroub, dental, chest, stomach, kidney, bladder, diabetes, feet, fracture, studyage, sex}

3.2 Discretization by Equal Frequency Binning

Equal frequency binning [1] technique considers each condition attribute individually. The discretization method divides the data into a specified number of intervals so that approximately the same number of objects fall into each of the intervals.

Table 6 lists the result of discretization using equal frequency discretization algorithm in ROSETTA. The first column shows the different of number of bins. The second column shows the number of inconsistent data after discretization. The third column lists the number of core attributes generated from data after removing the inconsistent data from the discretization. The fourth column gives the number of reduct sets generated by genetic algorithm. And the last column shows whether the core attributes are all contained in the reduct sets.

The experiments show that equal frequency discretization with 25 bins provides a best result. The core attributes generated from this discretization method are exactly the same as the core attributes from the original data set as shown in Table 4. For detailed result on equal frequency discretization with different number of bins, please refer to Appendix A.

Table 6. Core and reducts for the medical data set after Equal Frequency Binning discretization

Bin	Number of inconsistent data	Core Attributes	Reduct Sets by Genetic Algorithm	Contain Core
3	48	29	1	Yes
17	0	17	32	Yes
20	0	15	72	Yes
25	0	14	90	Yes

3.3 Discretization by Naive Algorithm

Naive algorithm takes both condition attributes and decision attributes into consideration [1]. The algorithm sorts the condition attributes first, then considers cut between two values of each attribute.

We also tried the naive discretization provided by ROSETTA. There is no inconsistent data after executing naive discretization algorithm. Table 7 shows the 14 core attributes. Table 8 shows the reduct sets generated by ROSETTA.

Both Johnson's algorithm and genetic algorithm return reduct sets that contains the core attributes. Genetic algorithm generates 66 reduct sets. The core attributes generated are the same as generated from the original data removing the inconsistent data as shown in Table 4.

Table 7. Core attributes for the medical data set after Naive discretization

Core	Core Attributes
Core	eartroub,livealone,heart,hbp,eyetroub, hearing,sex,health,edulevel,chest, housewk,diabetes,dental,studyage

4 Association Rules Generation

We can consider this medical data set as a transaction data set. We create an itemset with the patient's symptoms related to the attributes in the reduct and the decision attributes. For instance, if the attributes in the reduct are {*body temperature, blood pressure*} and the decision attribute is {*surviving or dead*}, an item set from a patient could be {*temperature_low, pressure_high, dead*}. Association rule algorithms can therefore be applied on this transaction data set to generate rules, which have condition attributes on the antecedent part and decision attributes on the consequent part of the rules. We perform association rules generation [4] based on the reduct set from Johnson's algorithm in Table 5

Table 8. Reduct sets for the medical data set after Naive discretization

Reduct Gen. Algorithm	Reduct Set
Johnson's Algorithm	{edulevel, eyesight, hearing, housewk, health, trouble, livealone, hbp, heart, arthriti, eyetroub, eartroub, dental, chest, bladder, diabetes, feet, nerves, studyage, sex}
Genetic Algorithm No.1	{edulevel, eyesight, hearing, walkout, housewk, health, trouble, livealone, cough, tired, sneeze, hbp, heart, arthriti, eyetroub, eartroub, dental, chest, bladder, diabetes, feet, nerves, studyage, sex}
No.2	{edulevel, eyesight, hearing, walkout, housewk, health, livealone, cough, tired, sneeze, hbp, heart, arthriti, eyetroub, eartroub, dental, chest, stomach, bladder, diabetes, nerves, fracture, studyage, sex}
...	...
No. 66	{edulevel, eyesight, hearing, walkout, meal, housewk, takemed, health, trouble, livealone, cough, sneeze, hbp, heart, arthriti, eyetroub, eartroub, dental, chest, stomach, kidney, bladder, diabetes, feet, nerves, skin, fracture, studyage, sex}

on Sun Fire V880, four 900Mhz UltraSPARC III processors, with 8GB of main memory. The rules are generated with $support = 30\%, confidence = 80\%$, as shown in Table 9. The subsumed rules are removed.

There are 16 rules generated by Johnson's algorithm. Noticed that the first 8 rules contain core attributes and only core attributes on the left side of the rules.

5 Summary

The experimental results show that data preprocessing on removing inconsistent data is important for generating core and reducts. The genetic algorithm provided by ROSETTA can be used to generate multiple reducts. And all the reducts contain the core attributes. Some rules generated from Johnson's reduct contain core attributes. This is an exciting discovery. We are interested in studying the effects of core attributes on improving the utility of association rules. We plan to apply association rules generation to the reducts from genetic algorithm in order to study our rule importance measure.

Table 9. The rules generated by Johnson's algorithm for the medical data set

No.	Rules
1	SeriousChestProblem → *Dead*
2	SeriousHearingProblem, HavingDiabetes → *Dead*
3	SeriousEarTrouble → *Dead*
4	SeriousEyeTrouble → *Dead*
5	SeriousHeartProblem → *Dead*
6	Livealone, HavingDiabetes, HighBloodPressure → *Dead*
7	VerySeriousHouseWorkProblem → *Dead*
8	Sex_2 → *Dead*
9	FeetProblem → *Dead*
10	SeriousEyeSight → *Dead*
11	Livealone, HavingDiabetes, NerveProblem, → *Dead*
12	TroublewithLife → *Dead*
13	LostControlofBladder, HavingDiabetes → *Dead*
14	Livealone, HighBloodPressure, LostControlofBladder → *Dead*
15	HighBloodPressure, LostControlofBladder, NerveProblem → *Dead*
16	Livealone, LostControlofBladder, NerveProblem → *Dead*

Acknowledgement

We would like to thank Dr. Arnold Mitnitski from Department of Medicine, Dalhousie University, for providing this Geriatric Care data set. Many thanks to Dr. Xiaohua Tony Hu from College of Information Science and Technology, Drexel University for valuable comments to the experiments.

References

1. Aleksander Ohrn:Discernibility and Rough Sets in Medicine: Tools and Applications. PhD Thesis, Department of Computer and Information Science, Norwegian University of Science and Technology, Trondheim, Norway, NTNU report 1999:133, IDI report 1999:14, ISBN 82-7984-014-1, 239 pages. 1999.
2. Aleksander Ohrn: ROSETTA Technical Reference Manual. Department ofComputer and Information Science, Norwegian University of Science and Technology, Trondheim, Norway. May 25, 2001.
3. Hu, X., Lin, T., Han, J.: A New Rough Sets Model Based on Database Systems. Fundamenta Informaticae XX(2004) 1-18.
4. Borgelt, C.: Efficient Implementations of Apriori and Eclat. Proceedings of the FIMI'03 Workshop on Frequent Itemset Mining Implementations. CEUR Workshop Proceedings (2003) 1613-0073
5. Pawlak, Z.: Rough Sets. In Theoretical Aspects of Reasoning about Data. Kluwer, Netherlands, 1991.

A Result on Discretization by Equal Frequency

A.1 Discretization by Equal Frequency Binning, Bin =3

After equal frequency discretization bin equal to 3, 48 of the 8535 data records are removed because of inconsistency. There are 8487 records left.

Table 10 shows the 29 core attributes. Table 11 shows the reduct sets generated by ROSETTA.

Table 10. Core attributes for the medical data set after Equal Frequency Binning discretization Bin=3

Core	Core Attributes
Core	eartroub,sneeze,meal,eyetroub,stomach, edulevel,kidney,chest,housewk,arthriti,tired, studyage,skin,feet,eyesight,trouble,livealone,hbp, heart,stroke,hearing,sex,health,nerves,diabetes, bladder,cough,dental,age6

Table 11. Reduct sets for the medical data set after Equal Frequency discretization Bin=3

Reduct Gen. Algorithm	Reduct Set
Johnson's Algorithm	{edulevel, eyesight, hearing, meal, housewk, health, trouble, livealone, cough, tired, sneeze, hbp, heart, stroke, arthriti, eyetroub, eartroub, dental, chest, stomach, kidney, bladder, diabetes, feet, nerves, skin, age6, studyage, sex}
Genetic Algorithm	{edulevel, eyesight, hearing, meal, housewk, health, trouble, livealone, cough, tired, sneeze, hbp, heart, stroke, arthriti, eyetroub, eartroub, dental, chest, stomach, kidney, bladder, diabetes, feet, nerves, skin, age6, studyage, sex}

The reduct contains exactly the 29 core attributes. But there is only one reduct generated.

A.2 Discretization by Equal Frequency Binning, Bin = 17

After equal frequency discretization bin equal to 17, there is no inconsistent data. So the data contains 8535 records.

Table 12 shows the 17 core attributes. Table 13 shows the reduct sets generated by ROSETTA. Genetic Algorithm returns 32 reduct sets.

Table 12. Core attributes for the medical data set after Equal Frequency Binning discretization Bin=17

Core	Core Attributes
Core	eartroub,eyetroub,stomach,edulevel,chest, housewk,studyage,feet,livealone,hbp, heart,hearing,sex,health,diabetes,dental,bladder

Table 13. Reduct sets for the medical data set after Equal Frequency Binning discretization Bin=17

Reduct Gen. Algorithm	Reduct Set
Johnson's Algorithm	{edulevel, eyesight, hearing, housewk, health, trouble, livealone, sneeze, hbp, heart, arthriti, eyetroub, eartroub, dental, chest, stomach, bladder, diabetes, feet, nerves, studyage, sex}
Genetic Algorithm No. 1	{edulevel, eyesight, hearing, housewk, health, trouble, livealone, sneeze, hbp, heart, arthriti, eyetroub, eartroub, dental, chest, stomach, bladder, diabetes, feet, nerves, studyage, sex}
No. 2	{edulevel, eyesight, hearing, meal, housewk, health, livealone, cough, sneeze, hbp, heart, arthriti, eyetroub, eartroub, dental, chest, stomach, kidney, bladder, diabetes, feet, nerves, age6, studyage, sex}
...	...
No. 32	{edulevel, hearing, shopping, housewk, health, trouble, livealone, cough, tired, sneeze, hbp, heart, stroke, arthriti, eyetroub, eartroub, dental, chest, stomach, kidney, bladder, diabetes, feet, nerves, skin, fracture, age6, studyage, sex}

A.3 Discretization by Equal Frequency Binning, Bin = 20

After equal frequency discretization bin equal to 20, there is no inconsistent data. So the data has 8535 records.

Table 14 shows the 15 core attributes. Table 15 shows the reduct sets generated by ROSETTA. Genetic Algorithm returns 72 reduct sets.

A.4 Discretization by Equal Frequency Binning, Bin = 25

After equal frequency discretization bin equal to 25, there is no inconsistent data. So the data has 8535 records.

Table 16 shows the 14 core attributes. And these 14 core attributes are the same as the core attributes in Table 4 generated from the original data after removing inconsistent data.

Table 14. Core attributes for the medical data set after Equal Frequency Binning discretization Bin=20

Core	Core Attributes
Core	eartroub,livealone,heart,hbp,eyetroub, hearing,stomach,sex,health,edulevel, chest,housewk,diabetes,dental,studyage

Table 15. Reduct sets for the medical data set after Equal Frequency Binning discretization Bin=20

Reduct Gen. Algorithm	Reduct Set
Johnson's Algorithm	{edulevel, eyesight, hearing, housewk, health, trouble, livealone, tired, hbp, heart, arthriti, eyetroub, eartroub, dental, chest, stomach, bladder, diabetes, nerves, studyage, sex}
Genetic Algorithm No. 1	{edulevel, eyesight, hearing, meal, housewk, money, health, trouble, livealone, sneeze, hbp, heart, arthriti, eyetroub, eartroub, dental, chest, stomach, kidney, bladder, diabetes, nerves, studyage, sex}
No. 2	{edulevel, eyesight, hearing, housewk, health, trouble, livealone, sneeze, hbp, heart, arthriti, eyetroub, eartroub, dental, chest, stomach, kidney, bladder, diabetes, nerves, skin, fracture, studyage, sex}
...	...
No. 72	{edulevel, eyesight, hearing, walkout, shopping, meal, housewk, takemed, health, trouble, livealone, cough, tired, sneeze, hbp, heart, eyetroub, eartroub, dental, chest, stomach, kidney, bladder, diabetes, feet, nerves, skin, studyage, sex}

Table 17 shows the reduct sets generated by ROSETTA. Genetic Algorithm returns 90 reduct sets.

B Attributes in the Original Data Set

Table 18 lists the attributes for this geriatric care data set.

C Background work on Rough Sets Theory

Based on Pawlak's book [5], we explain the basic concepts in rough sets theory.

Suppose we can represent a data set as a decision table, which is used to specify what kinds of conditions lead to the kinds of decisions. A decision table

Table 16. Core attributes for the medical data set after Equal Frequency Binning discretization Bin=25

Core	Core Attributes
Core	eartroub,livealone,heart,hbp,eyetroub, hearing,sex,health,edulevel,chest, housewk,diabetes,dental,studyage

Table 17. Reduct sets for the medical data set after Equal Frequency Binning discretization Bin=25

Reduct Gen. Algorithm	Reduct Set
Johnson's Algorithm	{edulevel, eyesight, hearing, housewk, health, trouble, livealone, hbp, heart, arthriti, eyetroub, eartroub, dental, chest, bladder, diabetes, feet, nerves, studyage, sex}
Genetic Algorithm No. 1	{edulevel, eyesight, hearing, housewk, money, health, trouble, livealone, cough, tired, hbp, heart, arthriti, eyetroub, eartroub, dental, chest, kidney, diabetes, feet, nerves, skin, fracture, studyage, sex}
No. 2	{edulevel, eyesight, hearing, housewk, health, trouble, livealone, cough, tired, hbp, heart, arthriti, eyetroub, eartroub, dental, chest, stomach, kidney, diabetes, feet, nerves, skin, fracture, studyage, sex}
...	...
No. 90	{edulevel, eyesight, hearing, walkout, shopping, housewk, takemed, money, health, trouble, livealone, cough, tired, sneeze, hbp, heart, stroke, eyetroub, eartroub, dental, chest, stomach, kidney, bladder, diabetes, feet, nerves, fracture, studyage, sex}

can be defined as $T = (U, C, D)$, where U is the set of objects in the table, C is the set of the condition attributes and D is the set of the decision attributes. Table 19 gives an example of the decision table. {a,b,c} is the set of condition attributes, and {d} is the set of decision attributes.

Here we only look at the situation when the value of the decision attributes is either 0, or 1. And we will not discuss the situation that the condition attributes have missing values.

U is the set of objects we are interested in, where $U \neq \phi$. Let R be an equivalence relation over U, then the family of all equivalence classes of R is represented by U/R. $[x]_R$ means a category in R containing an element $x \in U$. Suppose $P \subseteq R$, and $P \neq \phi$, $IND(P)$ is an equivalence relation over U. For any $x \in U$, the equivalence class of x of the relation $IND(P)$ is denoted as $[x]_P$. X

Table 18. Attributes for the geriatric care data set

Order	Name	Question
1	edulevel	Education level
2	eyesight	How is your eyesight?
3	hearing	How is your hearing?
4	eat	Can you eat?
5	dress	Can you dress and undress yourself?
6	takecare	Can you take care of your appearance?
7	walk	Can you walk?
8	getbed	Can you get in and out of bed?
9	shower	Can you take a bath or shower?
10	bathroom	Can you go to the bathroom commode?
11	phoneuse	Can you use the telephone?
12	walkout	Can you get places out of walking dist.?
13	shopping	Can you go shopping for groceries etc.?
14	meal	Can you prepare your own meals?
15	housewk	Can you do your housework?
16	takemed	Can you take your own medicine?
17	money	Can you handle your own money?
18	health	How is your health these days?
19	trouble	Trouble with life?
20	livealone	Do you live here alone?
21	cough	Often cough?
22	tired	Easy feel tired?
23	sneeze	Often sneeze?
24	hbp	High blood pressure?
25	heart	Heart problem?
26	stroke	Stroke or effects of stroke?
27	arthriti	Arthritis or rheumatism?
28	parkinso	Parkinson's disease?
29	eyetroub	Eye trouble not relieved by glasses?
30	eartroub	Ear trouble?
31	dental	Dental Problems?
32	chest	Chest problems?
33	stomach	Stomach or digestive problems?
34	kidney	Kidney Problems?
35	bladder	Lose control of your bladder?
36	bowels	Lose control of you bowels?
37	diabetes	Ever been diagnosed with diabetes?
38	feet	Feet problems?
39	nerves	Nerve problems?
40	skin	Skin problem?
41	fracture	Any fractures?
42	age6	Age group by 5-year
43	studyage	Age at investigation
44	sex	Sex
45	livedead	Survival status

Table 19. An Example of Decision Table

U	a	b	c	d
1	1	1	0	0
2	1	2	0	1
3	1	3	0	1
4	0	1	0	0
5	0	2	0	0
6	0	3	0	1
7	0	2	0	1
8	0	3	0	0

is a subset of U, R is an equivalence relation, the lower approximation of X and the upper approximation of X is defined as:

$$\underline{R}X = \bigcup \{x \in U | [x]_R \subseteq X\}$$

$$\overline{R}X = \bigcup \{x \in U | [x]_R \bigcap X \neq \phi\}$$

respectively.

Reduct and core are further defined as follows [5]. R is an equivalence relation and let $S \in R$. We say, S is dispensable in R, if $IND(R) = IND(R - \{S\})$; S is indispensable in R if $IND(R) \neq IND(R - \{S\})$. We say R is independent if each $S \in R$ is indispensable in R.

Q is a reduct of P if Q is independent, $Q \subseteq P$, and $IND(Q) = IND(P)$. An equivalence relation over a knowledge base can have many reducts. The intersection of all the reducts of an equivalence relation P is defined to be the *Core*, where

$$Core(P) = \bigcap \text{All Reducts of P.}$$

The reduct and the core are important concepts in rough sets theory. Reduct sets contain all the representative attributes from the original data set. It is often used in attribute selection process. The core is contained in all the reduct sets, and it is the necessity of the whole data. Any reduct generated from the original data set can not exclude the core attributes.

Let $T = (U, C, D)$ be a decision table, the C-positive region of D is defined to be the set of all objects of U which can be classified into U/D using attributes from C, which is,

$$POS_C(D) = \bigcup \{\underline{C}X | X \in IND(D)\}.$$

An attribute $f \in C$ is dispensable if $POS_{C-\{f\}}(D) = POS_C(D)$. All the core attributes are indispensable.

The degree of dependency between the equivalent class R and the decision attribute D is defined as

$$\tau_R(D) = \frac{\text{cardinality of } POS_R(D)}{\text{cardinality of } U}.$$

This dependency evaluation is often used as the stopping condition for the reduct generation algorithm.

Example 1. In Table 19, $U = \{1, 2, 3, ..., 8\}$ is a set of objects. $C = \{a, b, c\}$, $D = \{d\}$. Suppose $IND = \{b, c\}$. We have the equivalence classes of IND, $E_1 = \{1, 4\}$, $E_2 = \{2, 5, 7\}$, $E_3 = \{3, 6, 8\}$. The decision attribute d consists of two classes, $D_1 = \{2, 3, 6, 7\}$, $D_0 = \{1, 4, 5, 8\}$. The lower and upper approximation of D are,

$$\begin{aligned}
&\underline{R}D_1 = \phi \\
&\overline{R}D_1 = E_2 \cup E_3 = \{2, 3, 5, 6, 7, 8\} \\
&\underline{R}D_0 = E_1 = \{1, 4\} \\
&\overline{R}D_0 = E_1 \cup E_2 \cup E_3 = \{1, 2, 3, 4, 5, 6, 7, 8\}
\end{aligned}$$

Because $IND(\{b, c\}) = IND(\{b, c\} - \{c\})$, we say c is dispensable. For $P = \{a, b, c, d\}$, $Q \subseteq P$, $Q = \{a, b\}$. Because $IND(Q) = IND(P)$, $Q = \{a, b\}$ is a reduct of P.

$IND(D) = \{\{2, 3, 6, 7\}, \{1, 4, 5, 8\}\}$, $IND(\{b, c\}) = \{\{1, 4\}, \{2, 5, 7\}, \{3, 6, 8\}\}$, therefore $POS_{\{b,c\}}(D) = \{1, 4\}$.

Because $POS_{\{b,c\}-\{b\}}(D) = \phi \neq POS_{\{b,c\}}(D)$, b is indispensable.

$$\tau_{\{b,c\}}(D) = \frac{\text{cardinality of } POS_{\{b,c\}}(D)}{\text{cardinality of } U} = \frac{2}{4} = \frac{1}{2}.$$