RESEARCH NOTE

A NEW APPROACH FOR KNOWLEDGE-BASED SYSTEMS REDUCTION USING ROUGH SETS THEORY

G. A. Montazer

Department of Electrical Engineering, Institute of Information Technology
Tarbiat Modarres University, Tehran, 14115-179, Iran, montazer@modares.ac.ir

(Received: June 13, 2001 - Accepted in Final Form: February 6, 2003)

Abstract Problem of knowledge analysis for decision support system is the most difficult task of information systems. this paper presents a new approach based on notions of mathematical theory of Rough Sets to solve this problem. Using these concepts a systematic approach has been developed to reduce the size of decision database and extract reduced rules set from vague and uncertain data. The method has been applied to an imprical medical database with large scale data size and the final reduced and core rules has been extracted using concepts of this theory.

Key Words Information System, Database, Rough Sets Theory, Reduction, Decision Making

-Ç¼A nj /SwA »ÚBîÀÇ‡A ºB´ªTv¼w nj k®¼o- -½oT¦ñz« ,ºo¼£©¼ª~Ú ºB´ªTv¼w nj y¯Aj ¥¼¦dÚ **²k¼ña** »®TL« x°n -¼A/j±{»« ³ÀnA³§Fv« -¼A¥e º/oÙ™½›jß¯ ºBµ³î±ª\« »ƒB¼n·½oŠ¯ tBwA oÙ ºk½k] x°n³§BŠ« ©¼µB–Ç« •½oíÚBÙ°Bµ³~¼~i x pnA¤°k] ¥ñ{ ³Ù ²jAj ²B¢¼BQ·§BeA BÙ ¬Cnj ³Ÿ SwA »ñ¼ÚBªTv¼w nßñµh oÙ /j±{»Ç« Złohī wA S¼í†› ¨kî ° ¨B´ÙA BÙ ¨Ë±Ú¤°k] pA »~hz« kîA±›·î±ª\« ,¤°k] ·Īvµ ° ³T–B¼ yµBŸ¦ ºnBª¼Ù } ¼hzÚ¤°k]) »ñ{q» »í›A° ·§Fv« ð½ jßíÙA yµBŸ njk½k] x°n -¼A pA ²jß–ĪwA ¶±d¯ -¼A oÙ ²°Àî /SwA ²k¼jo£c½ozÚ ²k{ ²jBw kîA±› ¬jn°ℂ Swj³Ù »¢¯±¢a° ²k{ -¼¼LÚ (¬C©¼ÀÎ tBwA oÙ »æBi

1. INTRODUCTION

Information is often available in a form of databases known as information systems or attribute - value tables. The most difficult task of information systems is knowledge analysis for decision making process. Columns of an information table are labled by attributes, rows by objects and entries of the table are attribute values. Objects having the same attribute values are indiscernible with respect to these attributes [1].

Rough Sets Theory is a new approach to data analysis which has attracted attention of many researchers all over the world [2,3,4]. This theory overlaps with many other theories such as Fuzzy set theory [5], Evidence theory [6] and Boolean reasoning methods [7], nevertheless it can be viewed in its own rights, as an independent disipline [8].

This methodology has found many real - life applications in engineering [9, 10], medicine [11], image processing [12], and so on. The proposed approach has many advantages such as:

- Provides efficient algorithm for finding hidden pattern in data.
- Finds minimal Sets of data.
- Evalutes significance of data.
- It is Easy to understand and offers straightforward interpretation of results.

The main aspect of this article has focused on the core concepts of Rough Sets Theory and how to use it to reduce the size of the decision making processes to result in an evident and exact rule - based system from vague databases.

2. ROUGH SETS THEORY FUNDAMENTALS

The theory of rough sets has been under

continuous development for over recent years. The theory was originated by Zdzislaw Pawlak in 1970's as a result of a long term program of fundamental research on logical properties of information systems, carried out by him and a group of logicians from Polish Academy of Sciences and the University of Warsaw, Poland [13].

The methodology is concerned with the classificatory analysis of imprecise, uncertain or incomplete information or knowledge expressed in terms of data acquired from experience. The primary notions of the theory of rough sets are the approximation space and lower and upper approximations of a set. This process is called gronelling. The approximation space is a classification of the domain of interest into disjoint categories. The classification formally represents our knowledge about the domain, i.e. the knowledge is understood here as an ability to characterize all classes of the classification, for example, in terms of features of objects belonging to the domain. Objects belobging to the same category are not distinguishable, which means that their membership status with respect to an arbitrary subset of the domain may not always be clearly definable. This fact leads to the definition of a set in terms of lower and upper approximations. The lower approximation is a description of the domain objects which are known with certainty to belong to the subset of interest, whereas the upper approximation is description of the objects which possibly belong to the subset. Any subset defined through its lower and upper approximations is called "rough set".

The main specific problems addressed by the theory of rough sets are [8]:

1. Representation of uncertain or imprecise knowledge.
2. Empirical learning and knowledge acquisition from experience.
3. Knowledge analysis.
4. Analysis of conflicts.
5. Evaluation of the quality of the available information with respect to its consistency and the presence or absence of repetitive data patterns.
6. Identification and evaluation of data dependencies.
7. Approximate pattern classification.
8. Reasoning with uncertainty.
9. Information - preserving data reduction.

3. BASIC DEFINITIONS

Some basic definitions appear below, followed by a somewhat trivial example to demonstrate some basic concepts.

Given a set of objects, OBJ, a set of object attributes, AT, a set of values, VAL, and a function f: OBJ * AT → VAL, so that each object is described by the values of its attributes, we define an equivalence relation R(A), where A is a subset of AT: given two objects, o1 and o2; o1 R(A) o2 ⇔f(o1, a) = f(o2, a), for all a in A We say o1 and o2 are indiscernible (with respect to attributes in A). Now, we use this relation to partition the universe into equivalence classes, $\{e_0, e_1, e_2, ..., e_n\}$ = R*(A). The pair (OBJ, R) forms an approximation space with which we approximate arbitrary subsets of OBJ referred to as concepts.

Given O, an arbitrary subset of OBJ, we can approximate O by a union of equivalence classes:
The LOWER approximation of O (also known as the POSITIVE region):

LOWER (O) = POS (O) =
Union $\{e_i$ subset O$\}$

The UPPER approximation of O:

UPPER (O) = Union $\{e_i$ \ interset O \
not = empty$\}$

NEG (O) = OBJ - POS (O)

BND (O) = UPPER (O) - LOWER (O)

The latter definition is called BOUNDARY. The most common definition of a rough set is that a roughly definable set is a set, O, such that BND (O) is non-empty. So a rough set is a set defined only by its lower and upper approximations. A set, O, Whose boundary is empty, is exactly definable.

TABLE 1. Prototype Database.

Name	Education	Decision
Ali	High School	No
Maryam	High School	Yes
Hassan	Elementary	No
Hossein	University	Yes
Fatemeh	Doctorate	Yes

If a subset of attributes, A, is sufficient to create a partition R* (A) which exactly defines O, then we say that A is a reduct. The intersection of all reducts is known as the core. It must be noted that this is the simplest model and there are several probabilistic versions.

Often we use rough set theory for inductive learning description:

description [POS (O)] → Positive decision class

description [NEG (O)] → Negative decision class

description [BND (O)] →Probabilistically positive decision class

The simple following example reviews the meanings of the above concepts. Assume the above table which shows the relations among educational level and porospected job of five persons:

So, the set of positive examples of people with good job prospects:

O = {Maryam, Hossein, Fatemeh}

The set of attributes:

A = { Education }

The equivalence classes:

R*(A) = {(Ali,Maryam),(Hassan),(Hossein), (Fatemeh)}

The lower approximation and positive region:

POS (O) = LOWER (O) = {Hossein Fatemeh}

The negative region:

NEG (O) = {Hassan}

The boundary region:

BND (O) = {Ali, Maryam}

The Upper approximation:

UPPER (O) = POS (O) + BND (O) = {Ali, Maryam, Hossein, Fatemeh}

Using these definitions, decision rules will be derived as:

des [POS (O)] → Yes

des [Neg (O)] → No

des [BND (O)] → Possibly Yes or No

That is:

(Education, University) Or (Euducation, Doctorate) → Good prospects

(Education, Elementary) → No good prospect

(Education, High school) → Possible good prospect

4. PROBLEM STATEMENT

We have a diagnosis database which has compressed the expertness of especialists and experiment results about a special disease. This database such as shown in Table 2 has composed of five mesurement data A_1, ..., A_5, including the values of Ca^{2+}, $NaHPO_4$, P, K^+ and Fe^{3+}, respectively for 24 different cases which have been normalized to be in a comparable range H (High), M (Medium) and L (Low). These five parameters have been interpreted as attributes.

The classification of each state is made according to an expert and has classified by two possible outputs ‹Yes› and ‹No› which show if the case is under disease or not, respectively. One of the basic problems in this database is vagueness of decisions and uncertain relation between object - attribute value and its result (decision column). It is obvious that the larger size of the database, the more difficulties in decision process. Hence, we propose a method based on rough sets theory to reduce the size of foregoing information system and to classify the data entries to ease the decision making procedure.

5. ROUGH SET BASED REDUCTION APPROACH

The algorithm of the reduction of a decision table can be shown using algebric developments or based on logical relations [14]. Many algorithms have been developed to reduce the conditions and have been used in many

TABLE 2. Medical Decision Making Table.

Attributes					Decision
A_1	A_2	A_3	A_4	A_5	
L	H	M	H	H	Y
L	M	L	H	H	Y
M	H	H	H	L	Y
M	L	L	H	H	N
L	L	L	H	M	Y
M	M	L	H	L	Y
L	H	M	H	L	N
M	L	H	H	H	N
L	M	M	H	L	N
L	M	L	H	L	N
M	H	L	H	H	Y
L	M	M	H	H	N
L	M	M	M	L	Y
M	L	H	M	L	Y
L	H	M	L	H	N
M	M	H	L	M	Y
M	H	M	H	L	N
M	L	H	H	H	N
H	L	M	H	H	N
H	H	M	L	H	Y
H	L	L	M	H	Y
H	M	M	H	H	N
M	H	H	H	H	Y
L	L	L	H	M	Y

problems [7,2], but they have some difficulties which effect on their usefulness [11]. In this paper we present a modified procedure which collect the use and avoid the abuse of algorithm presented in References 7 and 2.

Basic steps in data analysis which can be tackled employing the rough set approach are the following:

- Characterization of set of objects in terms of attribute values;
- Finding dependenceies (total or partial) between attributes;
- Reduction of superfluous attributes (data);
- Decision rule generation.

This theory offers simple algorithms to conduct the above steps and enables straightforward interpretation of obtained results.

The first step of the algorithm is to verify if any attribute can be eliminated by repetition or not. In this database no attribute is similar for all of the samples. but there are some rows (samples or objects) which are identical, for example rows 5 and 24, 8 and 18. then resultant table has been shown in Table 3.

TABLE 3. Resultant Decision Table.

Attributes					Decision
A_1	A_2	A_3	A_4	A_5	
L	H	M	H	H	Y
L	M	L	H	H	Y
M	H	H	H	L	Y
L	L	L	H	M	Y
M	M	L	H	L	Y
M	H	L	H	H	Y
L	M	M	M	L	Y
M	L	H	M	L	Y
M	M	H	L	M	Y
H	H	M	L	H	Y
M	L	L	M	H	Y
M	H	H	H	H	Y
M	L	L	H	H	N
L	H	M	H	L	N
M	L	H	H	H	N
L	M	M	H	L	N
L	M	L	H	M	N
L	M	M	H	M	N
L	H	M	L	M	N
M	H	M	H	M	N
M	L	M	H	H	N
H	M	M	H	H	N

The next step is to verify if the decision table contains only indispensable attributes. This task can be accomplished eliminating step by step each attribute and verifying if the table gives the correct classification. For this aim, we must examine all of the samples whithout considering one of the attributes and find out whether or not the decision result changes and continue this method for all other attributes. Using this algorithm when the attribute A_1 is eliminated, we can verify that the rows 5 and 17 in Table 2 have the same entries but their decision results are different, so the attribute A_1 is indispensable. After following this step for all four other attributes, we can realize that eliminating attribute A_2, the rows 1 and 18 have the same difficulty. Hence, A_2 in not dispensable, too. But if the attributes A_3 and A_4 have been eliminated, then all of the data and results are compatible, therefor thses two attributes are

TABLE 4. Decision Table with Indispensable Attributes.

Attributes			Decision
A_1	A_2	A_5	
L	H	H	Y
L	M	H	Y
M	H	L	Y
L	L	M	Y
M	M	L	Y
M	H	H	Y
L	M	L	Y
M	L	L	Y
M	M	M	Y
H	H	H	Y
H	L	H	Y
M	H	H	Y
M	L	H	N
L	H	L	N
M	L	H	N
L	M	M	N
L	M	M	N
L	H	M	N
M	H	M	N
M	L	H	N
H	M	H	N

dispensable and can be removed. Finally, if the Attribute A_5 has been eliminated then the rows 1 and 14 have the same set of entries but different results. Table 4 presents this resultant set of objects.

Now, we can apply reduction mechanism to decrease the size of latter database. Using this approach the Table 5 has heen resulted and we can compute the core of the set of samples. This computation can be done eliminating each attribute step by step, and verifying if the decision table continues to be consistent. Applying this procedure for object 1, for instance, we can see that eliminating attribute A_1 preserves consistency in the table, and so does it for attribute A_2, but not for attribute A_3 (eliminating this attribute results inconsistence between rows 1 and 13). The results of applying this procedure for each objects of Table 4 has been represented in Table 6. This database shows the core of decision table and Table 7 contains the reduct of each object.

According to the latter table, knowledge existent in the Table 1 can be expressed by the following rules:

If $\{$(A_3 is H) or (A_3 is H and A_1 is L) or (A_1 is M and A_3 is L) or (A_2 is H and A_3 is H) or (A_3 is L) or (A_2 is M and A_3 is L) or (A_1 is M and A_2 is M) or (A_1 is H and A_2 is L)$\}$

then (Decision is Y)

If $\{$(A_1 is L and A_2 is H and A_3 is L) or (A_1 is L and A_2 is M and A_3 is M) or (A_2 is H and A_3 is M) or (A_1 is M and A_2 is L and A_3 is H) or (A_1 is H and A_2 is M)$\}$

then (Decision is N)

and using logical arithmetic we can express these rules by:

If $\{$(A_3 is H) or (A_3 is L) or (A_1 is M and A_2 is M) or (A_1 is H and A_2 is L)$\}$

then (Decision is Y)

If (A_2 is H) and $[$(A_1 is L and A_3 is L) or (A_3 is M)$]$ or (A_2 is M)

and $[$(A_1 is L and A_3 is M) or (A_1 is H)$]$ or (A_1 is M and A_2 is L and A_3 is H)

then (Decision is N)

which shows the explicit and essential rules for decision making.

TABLE 5. Decision Table with Reduced Objects.

Attributes			Decision
A_1	A_2	A_5	
L	H	H	Y
L	M	H	Y
M	H	L	Y
L	L	M	Y
M	M	L	Y
M	H	H	Y
L	M	L	Y
M	L	L	Y
M	M	M	Y
H	H	H	Y
H	L	H	Y
L	H	L	N
L	M	M	N
L	H	M	N
M	H	M	N
M	L	H	N
H	M	H	N

TABLE 6. Core of Database.

Attributes			Decision
A_1	A_2	A_5	
—	—	H	Y
L	—	H	Y
M	—	L	Y
—	H	H	Y
—	—	L	Y
—	H	H	Y
—	M	L	Y
M	—	L	Y
M	M	—	Y
—	H	H	Y
H	L	—	Y
L	H	L	N
L	M	M	N
—	H	M	N
—	H	M	N
M	L	H	N
H	M	—	N

TABLE 7. Reduct of the Database.

Attributes			Decision
A_1	A_2	A_3	
*	*	H	Y
L	*	H	Y
M	*	L	Y
*	H	H	Y
*	*	L	Y
*	M	L	Y
M	M	*	Y
H	L	*	Y
L	H	L	N
L	M	M	N
*	H	M	N
M	L	H	N
H	M	*	N

6. CONCLUSION

Knowledge base is one of the most important parts of intelligent systems which contains information and expertness of especialists. The knowledge acquisition mechanism is the most difficult task during the construction of information system.

Rough set theory presents a method which applying it we can extract the main aspects of information and essential data, using the notions of gronelling of universe of discourse, core and reduct of data base.

In this paper a systematic approach has been propersed based on Rough Set Theory to trarsform vague data in a reduced set of rules. This method has been based on the logical concepts and uses arithmetic tools to reduce the size of database and results in core of knowledge.

7. REFERENCES

1. Skoworn, A. and C. Rauszer; "The Discernibility Matrices and Functions in Information System", Decision Support by Experience, Kluwer Academic Publishers, (1992), 331-362.
2. Modrzejewski, M, "Feature Selection Using Rough Sets Theory", Machine Learning: ECML-93, Springer Verlag, Berlin, (1993), 213-226.
3. Nowicki, R., Slowinski, R. and Stefanowski, J., "Evaluation of Vibroacoustic Diagnostic Symptoms by Means of the Rough Sets Theory", ***Computers in Industry***, Vol. 20, No. 2, (Aug. 1992), 141-152.
4. Walczak, B. and Massart, D. L., "Rough Sets Theory", ***Chemometrics and Intelligent Laboratory Systems***, Vol. 47, No. 1, (Apr. 1999), 1-16.
5. Pawlack, Z., "Hard and Soft Sets in Rough Sets, Fuzzy sets and Knowledge Discovery", Ed. W. P. Ziarko, Springer-Verlag, Berlin, (1994), 130-135.
6. Skowron, A. and Grzymala-Busse, J., "From Rough Set Theory to Evidence Theory", In advances in the Dampster-Shafer Theory of Evidence, John Wiley and Sons, N.Y., (1994), 193-235, 251-271.
7. Krusinska, E., et al., "Discriminant Versus Rough Set Approach to Vague Data Analysis", ***Journal of Applied Statistics and Data Analysis***, Vol. 8, (1992), 43-56.
8. Pawlak, Z., "Rough Sets, Theoritical Aspects of Reasoning about Data", Kluwer Academic Publishers, London, (1992).
9. Munakata, T., "Rough Control: A Perspective", ***Proc. of 23rd Annual CSC***, California, (1995), 431-438.
10. Lambert-Torres, G., et al, "Classification of Power System Operation Point Using Rough Set Techniques", ***Proc. of Canadian Conf. on Elec. and Comp. Eng.***, Calgary, (1996), 1898-1903.
11. Szladow, A. and Ziarko, W., "Rough Sets: Working with Imperfect Data", AI Expert, Vol. 8, (1993), 36-41.
12. Grzymala-Busse, J., "Rough Sets, Advances in Imaging and Electron Physics", ***J. of Image Proccessing*** (1996), 88-96.
13. Pawlak, Z., "Rough Sets", ***International Journal of Computer and Information Science***, Vol. 11, (1982), 341-356.
14. Lin, T. Y. and Lin, Q., "Rough Approximate Operators: Axiomatic Rough Set Theory", Edited by W. P. Ziarko, Springer-Verlag, London, (1994), 256-260.