Identification and Evaluation of Functional Dependency Analysis using Rough sets for Knowledge Discovery

Y.V.Sreevani[1], Prof. T. Venkat Narayana Rao[2]
Department of Computer Science and Engineering,
Hyderabad Institute of Technology and Management,
Hyderabad, A P, INDIA
s_vanikumar@yahoo.co.in[1], tvnrbobby@yahoo.com[2]

Abstract— **The process of data acquisition gained momentum due to the efficient representation of storage/retrieving systems. Due to the commercial and application value of these stored data, Database Management has become essential for the reasons like consistency and atomicity in giving birth to DBMS. The existing database management systems cannot provide the needed information when the data is not consistent. So knowledge discovery in databases and data mining has become popular for the above reasons. The non-trivial future expansion process can be classified as Knowledge Discovery. Knowledge Discovery process can be attempted by clustering tools. One of the upcoming tools for knowledge representation and knowledge acquisition process is based on the concept of Rough Sets. This paper explores inconsistencies in the existing databases by finding the functional dependencies extracting the required information or knowledge based on rough sets. It also discusses attribute reduction through core and reducts which helps in avoiding superfluous data. Here a method is suggested to solve this problem of data inconsistency based medical domain with a analysis.**

Keywords- Roughset;knowledge base; data mining; functional dependency; core knowledge.

I. Introduction

The process of acquiring features hidden in the data is the major objective of Data Mining. Organizing these features for utilizing in planning for better customer satisfaction and promoting the business is the focus of Knowledge Representation. For discovering knowledge in data bases [6][7], in other words, reverse engineering has been attempted using the concept of Rough Sets for finding functional dependencies. Different phases of knowledge discovery process can be used for attribute selection, attribute extraction, data reduction, decision rule generation and pattern extraction. The fundamental concepts have been explored here for getting the core knowledge in Binary Data bases. Rough Sets are applied not only for Knowledge Representation but they are also being applied to Pattern Classification, Decision Making, Switching Circuits, and Data Compression etc[1][9]. It is proposed to find out the degree of dependency using Rough Sets introduced by Pawlak [1]which is used for characterizing the given instance, extracting information. This helps to know all Functional Dependencies existing in the vast databases. Rough set theory provides a collection of methods for extracting previously unknown data dependencies or rules from relational databases or decision tables. Rough set approach does not need any preliminary or additional information about data like probability in statistics, grade of membership in the fuzzy set theory. It proves to be efficient because it has got tools and algorithms which are sufficient for finding hidden patterns in data. It allows in reducing original data, i.e. to find minimal sets of data with the same knowledge as in the original data. The first, pioneering paper on rough sets, written by Zdzisław Pawlak, was published by International Journal of Computer and Information Sciences in 1982.

II. THE ROUGHSETS REPRESENTATIONS

The roughest method is basically associated with the classification and analysis of imprecise, uncertain or incomplete information or knowledge expressed in terms of data acquired from the experience. The domain is a finite set of objects. The domain of interest can be classified into two disjoint sets. The classification is used to represent 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 belonging 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 full certainty which undoubtedly belongs to the subset of interest, whereas the upper approximation is a description of the objects which would possibly belong to the subset. Any subset defined through its lower and upper approximations is called a rough set. The idea of rough set was proposed by Pawlak (1982) as a new mathematical tool to deal with vague concepts. Comer, Grzymala-Busse, Iwinski, Nieminen, Novotny, Pawlak, Obtulowicz, and Pomykala have studied algebraic properties of rough sets. Different algebraic semantics have been developed by P. Pagliani, I. Duntsch, M. K. Chakraborty, M. Banerjee and A. Mani; these have been extended to more generalize rough sets by D. Cattaneo and A. Mani, in particular. Rough sets can be used to represent ambiguity, vagueness and general uncertainty.

A. Knowledge Base

Let us consider a finite set U≠Ø (the universe) of objects under question, and R is a family of equivalence relations over U. Any subset $Q \subseteq U$ of the universe will be called a concept or a category in U and any family of concepts in U will be referred to as abstract knowledge (or in short knowledge) about U. A family of classifications over U will be called a knowledge base K over U. To this end we can understand knowledge base as a relational system K = (U, R), where U≠Ø is a finite set called the universe, and R is a family of equivalence relations over U. If E is an equivalence relation over U, then by E/R we mean the family of all equivalence classes of R (or classification of U) referred to as categories or concepts of R and $[Q]_R$ denotes a category in R containing an element q Є U. and If $P \subseteq R$ and $P \neq Ø$, then ∩ P (intersection of all equivalence relations belonging to P) is also an equivalence relation, and will be denoted by IND(P), and will be called an indiscernibility relation over P.

Therefore $[Q]_{IND(P)} = \cap\ [Q]_R$.

B. The Concept of Rough Sets

Let there be a relational system K = (U, **R**), where **U**≠Ø is a finite set called the universe, and **R** is a family of equivalence relations over **U**. Let $Q \subseteq \mathbf{U}$ and R be an equivalence relation [1]. We will say that Q is R-definable [12][1], if Q can be expressed as the union of some R-basic categories, otherwise Q is R-undefinable. The R-definable sets are called as R-exact sets some categories (subsets of objects) cannot be expressed exactly by employing available knowledge. Hence we arrive at the idea of approximation of a set by other sets. Let $Q \subseteq \mathbf{U}$ and equivalence relation RЄ IND(K) we associate two subsets i.e. $R_{LQ} = U\{Y Є \mathbf{U}/R : Y \subseteq Q\}$ and

$R^{UQ} = U\{Y Є \mathbf{U}/R : Y \cap Q \neq Ø\}$ called the R^{UQ} –UPPER and R_{LQ} -LOWER approximation of Q respectively[1][4].

From the above we shall also get the following denotations i.e.

$POS_R(Q) = R_{LQ}$, R-positive region of Q.

$NEG_R(Q) = \mathbf{U} - R^{UQ}$, R-negative region of Q.

$BN_R(Q) = R_{LQ} - R^{UQ}$,R-borderline region of Q.

The positive region $POS_R(Q)$ or the lower approximation of Q is the collection of those objects which can be classified with full certainty as members of the set Q, using Knowledge R.

In addition to the above we can define following terms-R-positive region of Q,$POS_R(Q) = R_{LQ}$.

Let $X \subseteq \mathbf{U}$, Where X is a subset of objects chosen from U and P and Q be the equivalence relations over U, then R-positive region of Q is $POS_P(Q) = \cup\, x \in u/_Q\ P_{LX}$

The P-positive region of Q is the set of all objects of the universe U which can be properly classified to classes of **U**/Q employing knowledge expressed by the classification **U**/P.

In the discovery of knowledge from huge databases we have to find the degree of dependency. This is used for characterizing the given instance, extracting information and which helps to know all the functional dependencies. Intuitively, a set of attributes Q depends totally on a set of attributes P, denoted P →Q, if the values of attributes from P uniquely determine the values of attributes from Q. In other words, Q depends totally on P, if there exists a functional dependency between values of P and Q. $POS_P(Q) = \cup\, x \in u/_Q\ P_{LX}$ called a positive region of the partition U/Q with respect to P, is the set of all elements of U that can be uniquely classified to blocks of the partition U/Q, by means of P. The degree of dependency between P and Q where $P,Q \subset \mathbf{R}$ is defined as follows.

If P and Q be a set of the equivalence relations over U,

Then the set of attributes of Q depends in a degree k ($0 \leq k \leq 1$), from P denoted by P → Q ,
If $k = \boldsymbol{\gamma}_P(Q)$ = Card $POS_P(Q)$/ C ard **U.**

Where card denotes cardinality of the Set and the symbol γ is used to specify POS that is positive region.

If k=1, we will say that Q totally depends from P.

If O<k<1, we say that Q partially depends from P.

If k=0, we say that Q is totally independent from P.

If k = 1 we say that Q depends totally on P, and if k < 1, we say that Q depends partially (to degree k) on P.
If k = 0 then the positive region of the partition U/Q with respect to P is empty.
The coefficient k expresses the ratio of all elements of the universe, which can be properly classified to blocks of the partition U/Q, employing attributes P and will be called the degree of the dependency. Q is totally (partially) dependent on P, if all (some) elements of the universe U can be uniquely classified to blocks of the partition U/Q, employing P. If the positive region is more then ,there exists a larger dependency between P and Q. This can be used to find the dependency between attribute sets in databases.
The above described ideas can also be interpreted as an ability to classify objects. more clearly, if k=1, then all elements of the knowledge base can be classified to elementary categories of U/Q by using knowledge P. If k≠1, only those elements of the universe which belong to the positive region can be classified to categories of knowledge Q, employing knowledge P. In particular if k=0, none of the elements of the universe can be classified using P and to elementary categories of knowledge Q.
More presicely, from the definition of dependency follows, that if, then the positive region of partition U/Q induced by Q

covers k*100 percent of all objects in the knowledge base. On the other hand, only those objects belonging to positive region of the partition can be uniquely classified. This means that k*100 percent of objects can be classified into block of partition U/Q employing P. If we restrict the set of objects in the knowledge base $POS_P(Q)$,we would obtain the knowledge base in which $P \rightarrow Q$ is a total dependency.

C. Indiscernibility

The notion of indiscernibility is fundamental to rough set theory. Informally, two objects are indiscernible if one object cannot be distinguished from the other on the basis of a given set of attributes. Hence, indiscernibility is a function of the set of attributes under consideration. An indiscernibility relation partitions the set of facts related to a set of objects into a number of equivalence classes . An equivalence class of a particular object is simply the collection of those objects that are indiscernible to the object in question[8] [13]. It is often possible that some of the attributes or some of the attribute values are superfluous. This enables us to discard functionally redundant information. A reduct is defined as a minimal set of attributes that preserves the meaning of indiscernibility relation [9][10] computed on the basis of the full set of attributes. Preserving the indiscernibility preserves the equivalence classes and hence it provide us the ability to form approximations. In practical terms, reducts help us to construct smaller and simpler models, and provide us an idea on the decision-making process [6],[7]. Typically, a decision table may have many reducts. However, there are extended theories to rough sets where some of the requirements are lifted. Such extensions can handle missing values and deal with hierarchies among attribute values. . In the following, for the sake of simplicity, it will be assumed that none of the attribute values are missing in data table so as to make it easy to find the dependencies.

D. Reduct and core Pertaining to Condition Attributes

Reduct and core of condition attributes helps in removing of superfluous partitions (equivalence relations) or/and superfluous basic categories in the knowledge base in such a way that the set of elementary categories in the knowledge base is preserved. this procedure enables us to eliminate all unnecessary knowledge from the knowledge base and preserving only that part of the knowledge which is really useful[13][14].

This concept can be formulated by the following example as follows.

Let $F=\{X_{1...}X_N\}$ is a family of sets choosen from **U** such that $X_i \subseteq$ **U.**

We say that X_i is dispensable in F,if $\cap(F-\{X_i\}) = \cap F$.

The family F is independent if all of its components are indispensible in F; otherwise F is dependent.

The family $H \subseteq F$ is a reduct of F, if H is independent and $\cap H= \cap F$.

The family of all indispensable sets in F will be called as the core of F, denoted by CORE(F).

From the above theory available in Rough Sets proposed by Pawlak.Z(1995) [1] the following definition can be derived where $CORE(F)= \cap RED(F)$ and RED(F) is the family of all reducts of F.

For example Consider family **R** = {P,Q,R} of equivalence relations having the following equivalence classes :
$U/P = \{\{x_1,x_3 , x_4 , x_5, x_6, x_7 \}, \{x_2, x_8\}\}$
$U/Q = \{\{x_1,x_3 , x_4 , x_5\}, \{x_2,x_6, x_7,x_8\}\}$
$U/R = \{\{x_1, x_5, x_6\}, \{x_2, x_7, x_8\}, \{x_3, x_4\}\}$
The family **R** induces classification
U/IND (**R**) = $\{\{x_1, x_5\} \{x_3, x_4\}, \{x_2, x_8\} \{x_6\}, \{x_7\}\}$
Moreover, assume that the equivalence relation S is given with the equivalence classes $U/S = \{\{x_1, x_5, x_6\}, \{x_3, x_4\}, \{x_2, x_7\}, \{x_8\}\}$. The positive region of S with respect to R is the union of all equivalence classes of U/IND(**R**) which are included in some equivalence classes of U/S, i.e. the set $POS_{\mathbf{R}}(S) = \{x_1,x_3 , x_4 , x_5, x_6, x_7\}$.
In order to compute the core and reducts of **R** with respect to S, we have first to find out whether the family **R** is S-dependent or not. According to definitions given in this section, we have to compute first whether P, Q and R are dispensable or not with respect to S (S-dispensable).
Removing P we get U/IND(R-{P}) = {{x1, x 5} {x3 , x4}, {x 2, x7, x 8} {x 6}}
Because, POS(R-{P})(S) = {x1, x3, x4, x5,x 6}$\neq POS_R(S)$
the P is S-indispensable in **R.**

Dropping Q from **R** we get U/IND(**R**-{Q}) = $\{\{x_1, x_5, x_6\}, \{x_3, x_4\}, \{x_2, x_8\}, \{x_7\}\}$, which yields the positive region $POS_{(\mathbf{R}-\{Q\})}(S) = \{x_1,x_3 , x_4 , x_5, x_6, x_7\}= POS_R(S)$
Hence Q is S-dispensable in **R**. Finally omitting R in **R** we obtain U/IND(**R**-{R}) = $\{\{x_1, x_3, x_4, x_5\}, \{x_2, x_8\}, \{x_6, x_7\}\}$ and the positive region is : $POS_{(\mathbf{R}-\{R\})}(S) = \varnothing \neq POS_{\mathbf{R}}(S)$, which means that R is S-indispensable in **R**.
Thus the S-core of **R** is the set {P,R}, which is also the S-reduct of **R.**

III. PROPOSED METHOD

To find the dependency between any subset of attributes using rough sets we are using a decision table based on certain factors and circumstances related to the knowledge base or the domain we choose. Due to the inconsistent nature of the data [11], certain data values in the data table may be conflicting. Here, a method is suggested to solve this problem of data inconsistency based on the approach inspired by rough set theory by Pawlak.Z.(1995)[1]. Generate the powerset of condition attributes for each element in the powerset :

- Find the equivalence classes.
- Associate a decision attribute.
- Find the degree of dependency.

- Find the inconsistent objects where the attribute values of the decision attributes are different, even though the attribute values of condition attributes are same.
- Calculate the degree of dependency k. Display those objects whose degree of dependency lies between 0 and 1.Display the inconsistent objects set.
- End for
- End.

IV. CREATION OF DECISION TABLE FOR KNOWLEDGE DISCOVERY

Let there be a set X of interest and is unknown and we have only some information about it. Assuming some sets which are disjoint with X and some sets included in X so as to build good approximations to X and use them to reason out on X. In this paper we are considering an example of a group of individuals (Table 1) who are at a risk of influenza (Zdzislaw Pawlak,1995)[1].

TABLE I. Patient information table

	temp	cough	head_ ache	muscle _pain	influenza
p_1	normal	present	present	present	present
p_2	normal	present	absent	absent	present
p_3	medium	present	absent	absent	absent
p_4	medium	absent	present	present	present
p_5	medium	absent	present	present	absent
p_6	high	present	present	present	present
p_7	high	absent	present	present	present
p_8	high	absent	present	present	absent
p_9	high	absent	absent	absent	absent

F_1----temp (normal, 0) (medium, 1) (high, 2)
F_2---- cough (present, 1) (absent, 2)
F_3----head_ ache (present, 1)(absent, 2)
F_4-----muscle_ pain (present, 1) (absent, 2)
F_5----influenza (present, 1) (absent, 2)

V. DECISION RULES

A decision rule [1],[5] is defined to be a logical expression in the form .IF (condition …) then (decision…), where in the condition is a set of elementary conditions connected by "and" and the decision is a set of possible outcomes/actions connected by "or". The above mentioned decision rule can be interpreted within the rough set framework and the If- then-part of the rule lists more than one possible outcome, that can be interpreted as describing one or more cases[8] . The If-then-part of the rule lists a single action Yes (or No.), that can be interpreted for describing one or more cases that lie in either the inside (or the outside) region of the approximation [14]. A set of decision rules forms the decision algorithm. Based on the above theory we consider the physical conditions related to nine patients (Table 1 and Table II) and their corresponding characteristic attribute values are used to derive the following rules which in turn help in building the decision table more rational.

Rule 1: if (temp=normal and cough=present and
head_ache=present and muscle_pain=present)
then (Influenza=present) .
Rule 2:if (temp=normal and cough=present and
head_ache=absent and muscle_pain=absent)
then (Influenza=present) .
Rule 3:if(temp=medium and cough=present and
head_ache=absent and muscle_pain=absent)
then (Influenza=absent) .
Rule 4:if (temp=medium and cough=absent and
head_ache=present and muscle_pain=present)
then (Influenza=present) .
Rule 5:if (temp=medium and cough=absent and
head_ache=present and muscle_pain=present)
then (Influenza=absent) .
Rule 6:if (temp=highand cough=present and
head_ache=present and muscle_pain=present)
then (Influenza=present) .
Rule 7: if(temp=high and cough=absent and head_ache=
present and muscle_pain= present)
then (Influenza=present).
Rule 8: if(temp= high and cough= absent and
head_ache= present and muscle_pain= present)
then (Influenza= absent) .
Rule 9:if(temp= high and cough= absent and
head_ache= absent and muscle_pain= absent)
then (Influenza= absent).

Using the above rules we can construct a decision table (Table I.) for nine different patients having different characteristics who are at a risk of influenza. The columns are labeled by factors or circumstances that reflect the physical condition of the patient in terms of set of condition attributes and decision attributes. The rows are labeled by objects where in each row represents a piece of information about the corresponding to each patient. Once the relation/table is created it is possible to find all the functional dependencies (Table III.) which would be useful for decision support systems as well as knowledge building/rule generation.

TABLE II. Decision Table

U	*Condition attributes*				*Decision attribute*
	F_1	F_2	F_3	F_4	F_5
p_1	0	1	1	1	1
p_2	0	1	2	2	1
p_3	1	1	2	2	2
p_4	1	2	1	1	1
p_5	1	2	1	1	2
p_6	2	1	1	1	1
p_7	2	2	1	1	1
p_8	2	2	1	1	2
p_9	2	2	2	2	2

The power set generated for the above condition attributes are:
$\{F_1\}_,\{F_2\}_,\{F_3\}_,\{F_4\}$
$\{F_1,F_2\},\{F_1,F_3\},\{F_1,F_4\},\{F_2,F_3\},\{F_2,F_4\},\{F_3,F_4\}$
$\{F_1,F_2,F_3\},\{F_1,F_3,F_4\},\{F_2,F_3,F_4\}$,$\{F_1,F_2,F_3,F_4\}$.

Using the power set we can generate various attribute sets for which functional dependencies are to be identified i.e. from the above table.

The equivalence classes for each element for the powerset are generated as below.

$U/F_{1=} \{\{P_1,P_2\},\{P_3,P_4,P_5\},\{P_6,P_7,P_8,P_9\}\}$

$U/F_2=\{\{ P_1,P_2, P_3,P_6\},\{ P_4,P_5,P_7,P_8,P_9\}\}$

$U/F_3= \{\{P_1, P_4,P_5, P_6 ,P_7,P_8\}, \{ P_2, P_3,P_9\}\}$

$U/F_{4=} \{\{ P_1, P_4,P_5, P_6 ,P_7,P_8\}, \{ P_2, P_3,P_9\}\}$etc.

The equivalence classes for decision attribute are:

$U/F_{5=} \{\{P_1,P_2, P_4, P_6 ,P_7\}, \{ P_3, P_5,P_8,P_9\}\}$.

By applying the theory of rough sets , The equivalence classes generated between every element of powerset and the decision attribute F_5 are :

$F_1 \rightarrow F_5$

$F_1= \{\{P_1, P_2\}, \{P_3,P_4,P_5\}, \{P_6,P_7,P_8,P_9\}\}$

$F_5= \{\{P_1 ,P_2, P_4, P_6, P_7\}, \{P_3,P_5,P_8,P_9\}\}$.

POS_{F1} (F5) = $\{P_1, P_2\}$, k= γ_{F1} (F5) = 2/9.

$F_2 \rightarrow F_5$

$F_2=\{\{ P_1,P_2, P_3,P_6\}, \{ P_4,P_5,P_7,P_8,P_9\}\}$.

$F_5=\{\{P_1, P_2, P_4, P_6 , P_7\}, \{ P_3, P_5,P_8,P_9\}\}$.

POS_{F2} (F5) = {0} , k= γ_{F2} (F5) = 0/9.

$F_3 \rightarrow F_5$

$F_3= \{\{ P_1, P_4, P_5, P_6 , P_7, P_8\}, \{ P_2, P_3, P_9\}\}$

$F_5= \{\{P_1, P_2, P_4, P_6, P_7\}, \{ P_3, P_5, P_8, P_9\}\}$.

POS_{F3} (F5) = {0} , k= γ_{F3} (F5) = 0/9.

$F_4 \rightarrow F_5$, $F_4= \{ P_1, P_4,P_5, P_6 ,P_7,P_8\}, \{ P_2, P_3,P_9\}$

$F_5= \{P_1,P_2, P_4, P_6 ,P_7\}, \{ P_3, P_5,P_8,P_9\}$, POS_{F4} (F5) = {0}.

k= γ_{F4} (F5) = 0/9 .

$F_1F_2 \rightarrow$ F5

$F_1F_2= \{\{P_1,P_2\}, \{P_3\}, \{P_4,P_5\}, \{P_6\}, \{P_7,P_8,P_9\}\}$

$F_5= \{\{P_1,P_2, P_4, P_6 ,P_7\}, \{ P_3, P_5,P_8,P_9\}\}$.

POS $_{F1F2}$ (F5) = $_=\{P_1,P_2, P_6 , P_3\}$. k= γ_{F1F2} (F5) = 4/9.

$F_1F_3 \rightarrow$ F5 ,

$F_1F= \{\{P_1\}, \{P_2\}, \{P_3\}, \{P_4,P_5\}, \{P_6,P_7,P_8\}, \{P_9\}\}$

$F_5=\{P_1, P_2, P_4, P_6 ,P_7\}, \{ P_3, P_5,P_8,P_9\}\}$.

POS $_{F1F3}$ (F5) = $\{P_1, P_2, P_3, P_9\}$. k= γ_{F1F3} (F5) = 4/9.

$F_1F_4 \rightarrow$ F5

$F_1F_4 = \{\{P_1\}, \{P_2\},\{P_3\}, \{P_4,P_5\}, \{P_6,P_7,P_8\}, \{P_9\}\}$

$F_5= \{\{P_1,P_2, P_4, P_6 ,P_7\}, \{ P_3, P_5,P_8,P_9\}\}$.

POS $_{F1F4}$ (F5) = $\{P_1,P_2, P_3 , P_9\}$. k= γ_{F1F4} (F5) = 4/9.

$F_2F_3 \rightarrow$ F5

$F_2F_3= \{\{P_1, P_6\}, \{ P_2, P_3\}, \{P_4,P_5,P_7,P_8\}, \{P_9\}\}$

$F_5= \{\{P_1, P_2, P_4, P_6 , P_7\}, \{ P_3, P_5,P_8,P_9\}\}$.

POS $_{F2F3}$ (F5) = $\{P_1, P_6, P_9\}$. k= γ_{F2F3} (F5) = 3/9.

$F_2F_4 \rightarrow$ F5

$F_2F_4= \{\{P_1,P_6\}, \{ P_2, P_3\}, \{P_4,P_5,P_7,P_8\}, \{P_9\}\}$.

$F_5= \{\{P_1,P_2, P_4, P_6 ,P_7\}, \{ P_3, P_5,P_8,P_9\}\}$.

POS $_{F2F4}$ (F5) = $\{P_1, P_6 , P_9\}$. k= γ_{F2F4} (F5) = 3/9.

$F_3F_4 \rightarrow$ F5

$F_3F_4= \{\{P1,P_4,P_5,P_6,P_7,P8\}, \{P_2,P_3,P_9\}\}$.

$F_5= \{\{P_1,P_2, P_4, P_6 ,P_7\}, \{ P_3, P_5,P_8,P_9\}\}$.

POS $_{F3F4}$ (F5) = {0}.

k= γ_{F2F4} (F5) = 0/9.

$F_1F_2F_3 \rightarrow F_5$

$F_1F_2F_3= \{\{P_1\}, \{P_2\},\{ P_3\}, \{P_4,P_5\}, \{P_6\}, \{P_7,P_8\}, \{P_9\}\}$.

$F_5= \{\{P_1, P_2, P_4, P_6 , P_7\}, \{ P_3, P_5,P_8,P_9\}\}$.

POS $_{F1F2F3}$ (F5) = $\{P_1, P_2, P_6, P_3 , P_9\}$. k= γ_{F1F2F3} (F5) = 5/9.

$F_2F_3F_4 \rightarrow F_5$

$F_2F_3F_4= \{\{P_1, P_6\}, \{P_2, P_3\}, \{P_4, P_5, P_7, P_8\}, \{P_9\}\}$.

$F_5= \{\{P_1, P_2, P_4, P_6, P_7\}, \{ P_3, P_5, P_8, P_9\}\}$.

POS $_{F2F3F4}$ (F5) = $\{P_1, P_6, P_9\}$. k= γ_{F2F3F4} (F5) = 3/9.

$F_1F_2F_4 \rightarrow F_5$

$F_1F_2F_4= \{\{P_1\}, \{P_2\}, P_3\}, \{P_4,P_5\}, \{P_6\}, \{P_7,P_8\}, \{P_9\}\}$.

$F_5= \{\{P_1, P_2, P_4, P_6 ,P_7\}, \{ P_3, P_5, P_8, P_9\}\}$.

POS $_{F1F2F4}$ (F5) = $\{P_1, P_2, P_6, P_3, P_9\}$. k= γ_{F1F2F4} (F5) = 5/9.

$F_1F_3F_4 \rightarrow F_5$

$F_1F_3F_4= \{\{P_1\},\{P_2\},\{ P_3\}, \{P_4,P_5\},\{P_6,P_7,P_8\},\{P_9\}\}$

$F_5= \{\{P_1,P_2, P_4, P_6 ,P_7\}, \{ P_3, P_5, P_8, P_9\}\}$.

POS $_{F1F3F4}$ (F5) = $\{P_1, P_2, P_3 , P_9\}$. k= γ_{F1F3F4} (F5) = 4/9.

$F_1F_2F_3F_4 \rightarrow F_5$

$F_1F_2F_3F_4= \{\{p_1\}, \{p_2\}, \{p_3\}, \{p_4,p_5\}, \{p_6\},\{ p_7,p_8\}, \{p_9\}\}$.

$F_5= \{\{P_1,P_2,P_4,P_6,P_7\}, \{P_3,P_5,P_8,P_9\}\}$.

POS $_{F1F2F3F4}$ (F5) = $\{P_1 , P_2, P_3 , P_6,P_9\}$. k= $\gamma_{F1F2F3F4}$ (F5) = 5 /9.

TABLE III. Dependency table

Power setelements(ps)	k= γ_{ps} (F_5)
F_1	2/9
F_2	0/9
F_3	0/9
F_4	0/9
F_1,F_2	4/9
F_1,F_3	4/9
F_1,F_4	4/9
F_2,F_3	3/9
F_2,F_4	3/9
F_3,F_4	0/9
F_1,F_2,F_3	5/9
F_1,F_3,F_4	4/9
F_2,F_3,F_4	3/9
F_1,F_2,F_4	5/9
F_1,F_2,F_3,F_4	5/9

VI. ANALYSIS BASED ON REDUCT AND CORE

By the above procedure we can extract Core of condition attributes which are explicitly necessary for deriving knowledge and coming to some conclusions related to the extraction of knowledge[2],[4]. We need to pursue a method which would give information of whether a particular characteristic attribute is necessary or not, based on which it can be established whether a patient has influenza or not. Analysis over the decision table is performed in this paper by identifying those core attributes whose removal would results in further inconsistency in the decision table which was consistent other wise. In the above decision table [2][3] (Table II.) by dropping F_1 rules P_2 and P_3 turns out to be inconsistent and positive region of the algorithm changes. Therefore, F_1 forms the core of the attribute set in the decision table. Similarly by dropping F_2 results in making P_6 and P_8 inconsistent and thus change in positive region of the algorithm[12]. The above procedure is repeatedly applied and checked for the inconsistency in the decision table and also to extract the core knowledge from the input domain.

U	*Condition attributes*			*Decision attribute*	
	***F*₁**	***F*₂**	***F*₃**	***F*₄**	***F*₅**
p₁	**0**	**1**	**1**	**1**	**1**
p₂	**0**	**1**	**2**	**2**	**1**
p₃	**1**	**1**	**2**	**2**	**2**
p₄	**1**	**2**	**1**	**1**	**1**
p₅	**1**	**2**	**1**	**1**	**2**
p₆	**2**	**1**	**1**	**1**	**1**
p₇	**2**	**2**	**1**	**1**	**1**
p₈	**2**	**2**	**1**	**1**	**2**
p₉	**2**	**2**	**2**	**2**	**2**

Figure 1. Attribute reduction

When we remove attribute F_1 rules 2 and 3 gets violated and the data corresponding to objects P_2 and P_3 turn into inconsistent as shown in Figure 1. But the removal of condition attribute F_4 still preserves the consistency of data and does not form the core of the condition attributes. The basic idea behind extraction of core knowledge is to retrieve knowledge of a characteristic attribute by observing its behavior, and this behavior is used to generate the algorithm and can be further used to simulate the actions in the future [4].

To find the reducts drop, take attributes as they appear in the power set and check whether any superfluous partitions (equivalence relations) or/and superfluous basic categories in the knowledge base [13] that are existing so that the set of elementary categories in the knowledge base is preserved. This procedure enables us to eliminate all unnecessary knowledge from the knowledge base, preserving only that part of the knowledge which is really useful.

For example drop F_4 and F_3 from $F_{1,}F_{2,}F_{3,}F_4$ and check the changes in the positive region.

This is done as follows

$$\text{Card(Pos}_{\{F1,F2,F3-\{F4\}\}})(F_5)= \{P_1,P_2,P_6,P_3,P_9\}=5$$
$$k= \gamma_P (Q) = 5/9$$
$$\text{Card(Pos}_{\{F1,F2,F4\}-\{F3\}} (F_5)= \{P_1,P_2,P_6,P_3,P_9\}=5$$
$$k= \gamma_P (Q) = 5/9.$$
$$\text{POS}_{\{\{F1,F2,F3 - \{F4\}\}}(F_5) = \text{POS}_{\{\{F1,F2,F4\}-\{F3\}\ \}}(F_5)=$$
$$\text{POS}_{F1,F2,F3,F4}\ (F_5)$$

From above we can notice that even with the removal of attributes F_4 and F_3 there is no change in the positive region.Therefore the reducts are {F1, F2, F3} and {F1, F2,F4} and the core attributes are { F1,F2}.

VII. CONCLUSION

Rough set theory provides a collection of methods for extracting previously unknown data dependencies or rules from relational databases or decision tables. As established above it can be said that roughsets relates to entities databases, data mining, machine learning, and approximate reasoning etc. This paper enables us to examine and to eliminate all unnecessary knowledge from the knowledge base by preserving only that part of the knowledge which is really useful. This paper gives some insight into roughsets which can be used to know data dependencies and extraction of knowledge. The ideas envisaged and depicted here are useful in the domain which deal huge collection of databases to analysis and take rational decisions in the areas such as banking, stock markets, medical diagnosis etc.

REFERENCES

[1] Zdzislaw Pawlak (1995)"Rough Sets: Theoretical Aspects Of Reasoning about Data", Institute of computer science, Noowowlejska 15/19, 00-665 Warsaw, Poland.

[2] H.Sug,"Applying rough sets to maintain data consistency for high degree relations", NCM'2008, Vol.22, 2008, pp.244-247.

[3] Duntsch, and G.Gediga,"Algebraic aspects of attribute dependencies in information systems", Fundamenta Informaticae, Vol.29,1997,pp.119-133.

[4] I. I.Duntsch, and G.Gediga,"Statistical evalution of rough set dependency analysis,"Internation journal of human computer studies,Vol.46,1997.

[5] T.Y.Lin, and H.Cao,"Seraching decision rules in very large databases using rough set theory,"Lecture notes in artificial intelligence Ziarco and Yao eds., 2000,pp.346-353.

[6] Arun K.Pujari,"Data Mining Techniques" Universities Press (India) Limited.

[7] Silbreschatz, Korth & Sudarshan (1997) "Database System Concepts"3/c McGraw Hill Companies, Inc.

[8] Dr.C.Raghavendra Rao, "Functional Dependencies and their role on Order Optimization.

[9] R.Stowinski, ed, Intelligent decision support: Handbook of Applications and advances of the rough set theory, Kulwer Academic publishers, 1992.

[10] A.Ohrn,Discernibility and rough sets in medicine: tools and Applications phd thesis,Department of computer and information science, Norwegian University of Science & Technology, 1999.

[11] J.G.Bazan, M.S.Szczuka, and j.Wroblewski,"A new Version of rough set exploration system,"Lecture notes in artificial intelligence,Vol. 2475, 2002,pp.397_404.

[12] N.Ttow, D.R.Morse, and D.M.Roberts,"Rough set approximation as formal concept,"journal of advanced computational intelligent Informatics, Vol.10, No.5, 2006, pp.606-611.

[13] Merzana kryszkiewicz,Piotr lasek**"**Fast Discovery of Minimal Sets of Attributes Functionally Determining a Decision Attribute" RSEISP '07**:** Proceedings of the international conference on Rough Sets and Intelligent Systems Paradigms .

[14] Hu.K, Sui.Y, Lu.Y, Wang.J, and Shi.C, "Concept approximation in concept lattice,*Proceedings of 5th Pacific-Asia Conference on Knowledge Discovery and Data Mining",PAKDD'01*, 167-173, 2001.

AUTHORS PROFILE

Graduated in AM.I.E.T.E. from I.E.T.E, New Delhi, India, in 1997 and M.Tech in Computer science from Osmania University,Hyderabad, A.P.,India in 2003. Currently working in Hyderabad Institute of Technology and Management as Associate professor in CSE department (HITAM) R.R.Dist, A.P, and India. She has 8 years of experience. Her research interests include Data mining, Distributed Systems, Information Retrieval systems.

Professor T.Venkat Narayana Rao, received B.E in Computer Technology and Engineering from Nagpur University, Nagpur, India, M.B.A (Systems) and M.Tech in Computer Science from Jawaharlal Nehru Technological University, Hyderabad, A.P., India and a Research Scholar in JNTU. He has 20 years of vast experience in Computer Science and Engineering areas pertaining to academics and industry related I.T issues. He is presently Professor and Head, Department of Computer Science and Engineering, Hyderabad Institute of Technology and Management (HITAM), Gowdavally, R.R.Dist., A.P, INDIA. He is nominated as an Editor and Reviewer to 15 International journals relating to Computer Science and Information Technology. He is currently working on research areas which include Digital Image Processing, Digital Watermarking, Data Mining, Network Security and other Emerging areas of Information Technology . He can be reached at tvnrbobby@yahoo.com