ROUGH SET APPROACH TO GENERATE CLASSIFICATION RULES FOR DIABETES

Sujogya Mishra[1], Shakti Prasad Mohanty[2], Sateesh Kumar Pradhan[3]
[1]Research scholar, Utkal University
Bhubaneswar-751004, India
[2]Department of Mathematics, College of Engineering and Technology
Bhubaneswar-751003, India
[3]Department of Computer Science, Utkal University
Bhubaneswar-751003, India

Abstract

In the current age medical science improved to certain height but some commonly disease like diabetes provide lots of symptoms , it is very tedious task to find the best fit symptom for diabetes to get accurate symptoms for diabetes we develop a technique using rough set concept which is précised and accurate up to certain extent. To start with, we take 100 samples and then using correlation techniques we consider 20 samples for our purpose, then apply rough set concept to find minimum number of symptoms for diabetes

Key words: Rough Set Theory, Medical related data, Granular computing, Data mining.

Introduction- The increasing size of data and the number of existing databases make difficult for humans to analyze it which creates both a need and an opportunity to ex-tract knowledge from databases[1] Medical databases have accumulated large quantities of information about patients and their medical conditions. Relationships and patterns within this data could provide new medical knowledge. Analysis of medical data is often concerned with treatment of incomplete knowledge, with management of inconsistent pieces of information and with manipulation of various levels of representation of data. Existing intelligent techniques[2] of data analysis are mainly based on quite strong assumptions knowledge about dependencies, probability distributions and large number of experiments are unable to de-rive conclusions from incomplete knowledge, or cannot manage inconsistent pieces of information. The standard intelligent techniques used in medical data analysis are neural network[3] Bayesian classifier [4] genetic algorithms[5] decision trees [6] fuzzy set [7] . Rough set theory and the basic concept was invented by Polish logician, Professor Z. Pawlak in early eighties[8] The theory of rough sets is a mathematical tool for extracting knowledge from un-certain and incomplete data based information. The theory assumes that we first have necessary information or knowledge of all the objects in the universe with which the objects can be divided into different groups. If we have exactly the same information of two objects then we say that they are indiscernible (similar), i.e., we cannot distinguish them with known knowledge. The theory of Rough Set can be used to find dependence relationship among data, evaluate the importance of attributes, discover the patterns of data, learn common decision-making rules, reduce all redundant objects and attributes and seek the minimum subset of attributes so as to attain satisfying classification. More-over, the rough set reduction algorithms enable to approximate the decision classes using possibly large and simplified patterns[9].This theory become very popular among scientists around the world and the rough set is now one of the most developing intelligent data analysis. Unlike other intelligent methods such as fuzzy set theory, Dempster–Shafer theory or statistical methods, rough set analysis requires no external parameters and uses only the information presented in the given data [10].This paper discusses how rough set theory can be used to analyze medical data, and for generating classification rules from a set of observed samples of the diabetes data. The rough set reduction technique is applied to find all reducts of the data which contains the minimal subset of attributes that are associated with a class label for classification. In our paper we organized it in to three section 1st section consists the introduction 2nd section consists the data analysis of the collected data 3rd section is the conclusion part, with a proper experimental section is mentioned before we conclude the paper .

.

2.PRILIMINARIES

2.1 Rough set Rough set theory as introduced by Z. Pawlak[8] is an extension of conventional set theory that support approximations in decision making.

2.1.2 Approximation Space: An Approximation space is a pair (U , R) where U is a non empty finite set called the universe R is an equivalence relation defined on U.

2.1.3 Information System: An information system is a pair S = (U , A), where U is thenon-empty finite set called the universe, A is the non-empty finite set of attributes

2.1.4 Decision Table: A decision table is a special case of information systems S= (U , A= C U {d}), where d is not in C. Attributes in C are called conditional attributes and d is a designated attribute called the decision attribute

.2.1.5 Approximations of Sets: Let S = (U, R) be an approximation space and X be a subset of U. The lower approximation of X by R in S is defined as $\underline{RX}$ = { e ε U | [e] ε X} and The upper approximation of X by R in S is defined as $\overline{RX}$□{*e*□*U*/[*e*]□*X* □□}where [e] denotes the equivalence class containing e. A subset X of U is said to be R-definable in S if and only if $\overline{RX}$=$\underline{RX}$ .A set X is rough in S if its boundary set is nonempty.

2.2 Dependency of Attributes

Let C and D be subsets of A. We say that D depends on C in a degree k (0 ≦k ≦1) denoted by C →k D if

K=y(C,D)=$\left|\dfrac{POS_C(D)}{|u|}\right|$ where POS$_C$(D) =U $\underline{C}$(X), is called positive region of the partition U/D with respect to C where $x \in u/d$, which is all elements of U that can be uniquely classified to the block of partition U/D. If k = 1 we say that D depends totally on C. If k < 1 we say that D depends partially (in a degree k) on C.

2.3 Dispensable and Indispensable Attributes- *Let S = (U, A = C υ D) be a decision table.* Let c be an attribute in C. Attribute c is dispensable in S if POS$_C$ (D)= POS$_{(C-\{c\})}$(D)otherwise, c is indispensable. A decision table S is independent if all attributes in C are indispensable. *Let S = (U, A = C υ D) be a decision table.*

Rough Set Attribute Reduction (RSAR) provides a filter based tool by which knowledge may be extracted from a domain in a concise way; retaining the information content whilst reducing the amount of knowledge involved.

2.4 Reduct and Core Let S = (U, A=C U D) be a decision table. A subset R of C is a reduct of C, if POS$_R$(D) = POS$_C$(D) and S' = (U, RUD) is independent, ie., all attributes in R are indispensible in S'. Core of C is the set of attributes shared by all reducts of C. CORE(C) = ∩RED(C) where, RED(C) is the set of all reducts of C. The reduct is often used in the attribute selection process to eliminate redundant attributes towards decision making.

2.5 **Correlation**- Correlation define as a mutual relationship or connection between two or more things .The quantity *r*, called the *linear correlation coefficient*, measures the strength and the direction of a linear relationship between two variables. The linear correlation coefficient is sometimes referred to as the *Pearson product moment correlation coefficient* in honor of its developer Karl Pearson. The mathematical formula for its coefficient given by the formula

$$r = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{\sqrt{n\left(\sum x^2\right) - \left(\sum x\right)^2}\sqrt{n\left(\sum y^2\right) - \left(\sum y\right)^2}}$$

2.6 **Goodness of fit**-The **goodness of fit** of a statistical model describes how well it fits a set of observations. Measures of **goodness of fit** typically summarize the discrepancy between observed values and the values expected under the model in question.

2.7 Chi **squared** distribution- A **chi-squared test**, also referred to as χ^2 **test**, is any statistical hypothesis test in which the sampling distribution of the test statistic is a chi squared distribution when the null hypothesis is true. Also considered a chi-squared test is a test in which this is *asymptotically* true, meaning that the sampling distribution (if the null hypothesis is true) can be made to approximate a chi-squared distribution as closely as desired by making the sample size large enough. The chi-square (I) test is used to determine whether there is a significant difference between the expected frequencies and the observed frequencies in one or more categories. Do the numbers of individuals or objects that fall in each category differ significantly from the number you would expect? Is this difference between the expected and observed due to sampling variation, or is it a real difference

2.7 **Further analysis of chi square test**- Basic properties of chi squared goodness fit is that it is non symmetric in nature .How ever if the degrees of freedom increased it appears to be to be more symmetrical .It is right tailed one sided test. All expectation in chi squared test is greater than 1.E$_I$=np$_i$ where n is the number samples considered p$_i$ is the probability of ith occurrence .Data selected at random there are two hypothesis null hypothesis and alternate hypothesis null hypothesis denoted by H$_0$ alternate hypothesis denoted by H$_1$. H$_0$ is the claim does follow the hypothesis and H$_1$ is the claim does not follow the hypothesis here H$_1$ is called the alternate hypothesis to H$_0$.If the test value found out to be K then K can be calculated by the formula K=∑(O$_I$-E$_I$)^2/ E$_I$. Choice of significance level always satisfies type 1 error .

2.8 **Different types of error**-

1) Type 1 error-Rejecting a hypothesis even though it is true 2) Type 2 error-Accepting the hypothesis when it is false

3) Type 3 error-Rejecting a hypothesis correctly for wrong reason

3. Basic idea

The basic idea for the proposed work is conceived from the general medical science. We initially consider 1000 samples, of Diabetes and five conditional attributes such as Sweating, Polyuria , Polydipsia , Polyphagia, restlessness . then, by correlation analysis , we consider 20 samples which are dissimilar in nature. Then we apply rough set concept to develop an algorithm, Which was appears to be précised, then we validate this by certain well known statistical validation method

4.Data Reduction

As the volume of data is increasing day by day, it is very difficult to find which attributes are important for a particular application and which are not that important and can be neglected. The aim of data reduction is to find the relevant attributes that have all essential information of the data set. The process is illustrated through the following 20 samples by using the rough set theory. In this particular problem we consider the conditional attributes sweating , Polyuria(need to urinate frequently) , Polydipsia(increased thirst & fluid intake) , Polyphagia(increased appetite), and restlessness as a$_1$,a$_2$,a$_3$,a$_4$,a$_5$ respectively and it's values are defined as moderate , severe , normal as b$_1$, b$_2$,b$_3$ respectively decision attributes are positive , negative as c$_1$,c$_2$ respectively. All the data collected from Dr Pradeep kumar mishra M.D.

Table-1:

E	a_1	a_2	a_3	a_4	a_5	d
E_1	b_2	b_2	b_1	b_1	b_1	c_1
E_2	b_2	b_2	b_1	b_3	b_3	c_1
E_3	b_1	b_2	b_2	b_3	b_3	c_2
E_4	b_1	b_2	b_2	b_3	b_3	c_1
E_5	b_3	b_3	b_3	b_3	b_2	c_2
E_6	b_1	b_2	b_2	b_2	b_2	c_1
E_7	b_2	b_2	b_2	b_2	b_2	c_1
E_8	b_1	b_1	b_1	b_1	b_1	c_2
E_9	b_1	b_2	b_2	b_3	b_3	c_1
E_{10}	b_1	b_2	b_2	b_2	b_2	c_2
E_{11}	b_2	b_3	b_3	b_3	b_3	c_2
E_{12}	b_1	b_2	b_3	b_1	b_2	c_1
E_{13}	b_3	b_2	b_2	b_2	b_1	c_2
E_{14}	b_3	b_3	b_3	b_3	b_3	c_2
E_{15}	b_2	b_1	b_1	b_1	b_1	c_1
E_{16}	b_1	b_1	b_1	b_1	b_1	c_1
E_{17}	b_1	b_3	b_2	b_2	b_3	c_2
E_{18}	b_1	b_2	b_2	b_3	b_2	c_1
E_{19}	b_1	b_3	b_1	b_3	b_1	c_2
E_{20}	b_2	b_2	b_2	b_3	b_3	c_1

The decision table -1 , takes the initial values before finding the reduct looking at the data table it is found that entities E_3,E_4, ambiguous in nature so both E_3,E_4 remove from the relational table -1 to produce the new table -2

Table-2:

E	a_1	a_2	a_3	a_4	a_5	d
E_1	b_2	b_2	b_1	b_1	b_1	c_1
E_2	b_2	b_2	b_1	b_3	b_3	c_1
E_5	b_2	b_1	b_2	b_3	b_2	c_2
E_6	b_1	b_2	b_2	b_2	b_2	c_1
E_7	b_2	b_2	b_2	b_2	b_2	c_1
E_8	b_1	b_1	b_1	b_1	b_1	c_1
E_9	b_1	b_2	b_2	b_3	b_3	c_1
E_{10}	b_1	b_2	b_2	b_2	b_2	c_2
E_{11}	b_2	b_2	b_1	b_3	b_3	c_2
E_{12}	b_1	b_2	b_1	b_1	b_2	c_1
E_{13}	b_1	b_2	b_2	b_2	b_1	c_2
E_{14}	b_2	b_2	b_2	b_3	b_3	c_2
E_{15}	b_2	b_1	b_1	b_1	b_1	c_1
E_{16}	b_1	b_1	b_1	b_1	b_1	c_1
E_{17}	b_1	b_2	b_2	b_2	b_3	c_2
E_{18}	b_1	b_2	b_2	b_3	b_2	c_2
E_{19}	b_1	b_2	b_1	b_3	b_1	c_2
E_{20}	b_2	b_1	b_2	b_3	b_3	c_1

Indiscernibility relation:

Indiscernibility Relation is the relation between two or more objects where all the values are identical in relation to a subset of considered attributes.

Approximation:

The starting point of rough set theory is the indiscernibility relation, generated by information concerning objects of interest. The indiscernibility relation is intended to express the fact that due to the lack of knowledge it is unable to discern some objects employing the available information Approximations is also other an important concept in Rough Sets Theory, being associated with the meaning of the approximations topological operations (Wu et al., 2004). The lower and the upper approximations of a set are interior and closure operations in a topology generated by the indiscernibility relation. Below is presented and described the types of approximations that are used in Rough Sets Theory.

Lower Approximation :

Lower Approximation is a description of the domain objects that are known with certainty to belong to the subset of interest.The Lower Approximation Set of a set X, with regard to R is the set of all objects, which can be classified with X regarding R, that is denoted as R_L.

a. Upper Approximation :

Upper Approximation is a description of the objects that possibly belong to the subset of interest. The Upper

Approximation Set of a set X regarding R is the set of all of objects which can be possibly classified with X regarding R . Denoted as R_U

b. <u>Boundary Region (BR) :</u>

Boundary Region is description of the objects that of a set X regarding R is the set of all the objects, which cannot be classified neither as X nor -X regarding R. If the boundary region X= ϕ then the set is considered "Crisp", that is, exact in relation to R; otherwise, if the boundary region is a set **X**≠ ϕ the set X "Rough" is considered. In that the boundary region is BR = R_U-R_L.

The lower and the upper approximations of a set are interior and closure operations in a topology generated by a indiscernibility relation. In discernibility according to decision attributes in this case has divided in to two groups

One group consist of all positive case and other one all negative cases

$E(Positive)=\{ E_1, E_2, E_6, E_7, E_8, E_9 ,E_{12}, E_{15}, E_{16},E_{20\backslash}\}$ ……….(1)

$E(Negative)=\{ E_5, E_{10}, E_{11}, E_{13}, E_{14} ,E_{17}, E_{18}, \}$……………….(2)

Here in this case lower approximation for positive cases represented by equation(1) and lower approximation for negative cases represented by equation(2)

Now find the entities which falls into a group as follows $E(sweating)_{moderate} =\{E_6, E_8, E_9, E_{10}, E_{12}, E_{16} ,E_{17}, E_{18}, E_{19}\}$ $E(sweating)_{severe} =\{E_1, E_2, E_7, E_{11} ,E_{15}, E_{20}\}$, $E(sweating)_{normal} =\{E_5, E_{13}, E_{14} \}$,

$E(polyuria)_{moderate} =\{E_8, E_{15}, E_{16} \}$, $E(polyuria)_{severe} =\{E_1, E_2, E_6,E_7,E_9,E_{10},E_{12},E_{13},E_{18},E_{20} \}$ $E(Polyuria)_{normal}=\{E_5,E_{11},E_{14},E_{17},E_{19}\}$ $E(Polydipsia)_{moderate} =\{ E_1, E_2, E_8,E_{15},E_{16},E_{19}\}$ $E(Polydipsia)_{severe} =\{ E_6,E_7,E_9,E_{10},E_{13},E_{17},E_{18},E_{20}\}$

$E(Polydipsia)_{normal} =\{ E_5,E_{11},E_{12},E_{14}\}$ $E(Polyphagia)_{moderate} =$

$\{ E_1,E_8,E_{12},E_{15}, E_{16} \}$

$E(Polyphagia)_{severe} =\{ E_6,E_7,E_{10}, E_{13} ,E_{17} \}$ $E(Polyphagia)_{normal} =\{ E_2,E_5,E_9,E_{11},E_{14},E_{18},E_{19},E_{20}\}$

$E(restlessness)_{moderate}=\{E_1,E_8,E_{13},E_{15},E_{16},E_{19}\}$ $E(restlessness)_{severe} =\{E_5,E_6,E_7,E_{10},E_{12},E_{18}\}$

$E(restlessness)_{normal} =\{E_2,E_9,E_{11},E_{14},E_{17},E_{20}\}$

Next, we find the combination of two attributes each to generate the reduct such combinations are $E(a_1,a_2)$, $E(a_1,a_3)$, $E(a_1,a_4)$, $E(a_1,a_5)$

$E(a_1,a_2)_{moderate}=\{E_8,E_{16}\}$, $E(a_1,a_2)_{severe}=\{E_1,E_2,E_7,E_{20}\}$, $E(a_1,a_2)_{normal} =\{E_3,E_{14} \}$ $E(a_1,a_3)_{moderate}=\{E_8,E_{16},E_{19}\}$ $E(a_1,a_3)_{severe} =\{E_7,E_{20}\}$

$E(a_1,a_3)_{normal} =\{E_5,E_{14}\}$ $E(a_1,a_4)_{moderate}=\{E_8,E_{12},E_{16}\}$ $E(a_1,a_4)_{severe} =\{E_7 \}$ $E(a_1,a_4)_{normal} =\{E_5,E_{14} \}$ $E(a_1,a_5)_{moderate}=\{E_8,E_{12},E_{16}\}$ $E(a_1,a_5)_{severe} =\{E_7\}$

$E(a_1,a_5)_{normal} =\{E_{14}\}$, $E(a_2 , a_3)_{moderate} =\{E_8,E_{15},E_{16}\}$ $E(a_2,,a_3)_{severe} =\{E_6,E_7,E_9,,E_{10},E_{13},E_{18},E_{20}\}$, $E(a_2,,a_3)_{normal} =\{E_5,E_{11},E_{14}\}$ $E(a_2,,a_4)_{moderate} =\{E_8,E_{15},E_{16}\}$

$E(a_2,,a_4)_{severe} =\{E_6,E_7,E_{10},E_{13}\}$ $E(a_2,,a_4)_{normal} =\{E_5,E_{11},E_{14} \}$ $E(a_2,,a_5)_{moderate} =\{E_8,E_{16} \}$ $E(a_2,,a_5)_{severe} =\{E_7\}$ $E(a_2,,a_5)_{normal} =\{E_{11},E_{14},E_{17}\}$

$E(a_3,,a_4)_{moderate} =\{E_1,E_8,E_{15},E_{16}\}$ $E(a_3,,a_4)_{severe} =\{E_6,E_7,E_{1\backslash 0},E_{13},E_{17}\}$ $E(a_3,,a_4)_{normal} =\{E_5, E_{11},E_{14}\}$ $E(a_3,,a_5)_{moderate} =\{E_1, E_8,E_{15},E_{16},E_{19}\}$

$E(a_3,,a_5)_{severe} =\{E_6, E_7,E_{10},E_{18} \}$ $E(a_3,,a_5)_{normal} =\{E_{11}, E_{14} \}$ $E(a_4,,a_5)_{moderate} =\{E_1, E_8 ,E_{15},E_{16}\}$ $E(a_4,,a_5)_{severe} =\{E_6, E_7 , E_{10} \}$ $E(a_4,,a_5)_{normal} =\{E_2,E_9,E_{20}\}$

$E(a_1,a_2 ,a_3)_{moderate}=\{E_8,E_{16}\}$ $E(a_1,a_2 ,a_3)_{severe} =\{E_7,E_{20}\}$ $E(a_1,a_2 ,a_3)_{normal} =\{E_5,E_{14}\}$ $E(a_2,a_3 ,a_4)_{moderate} =\{E_8 ,E_{15}, E_{16}\}$ $E(a_2,a_3 ,a_4)_{severe} =\{E_6 ,E_7, E_{10} ,E_{13}\}$

$E(a_2,a_3 ,a_4)_{normal} =\{E_5 ,E_{11}, E_{14} \}$ $E(a_3,a_4 ,a_5)_{moderate} =\{E_1,E_8, E_{15} \}$ $E(a_3,a_4 ,a_5)_{severe} =\{E_6,E_7, E_{10} \}$ $E(a_3,a_4 ,a_5)_{normal} =\{E_{11},E_{14}\}$

$E(a_1,a_2,a_3,a_4)_{moderate}=\{E_8,E_{16}\}$ $E(a_1,a_2,a_3,a_4)_{severe}=\{E_7\}$ $E(a_1,a_2,a_3,a_4)_{normal}=\{E_5,E_{14}\}$, these equivalent classes basically responsible for finding the dependencies with respect to the decision variable d in this paper besides all equivalence classes , we are trying to find out the degree of dependencies of different attributes of consideration with respect to decision attributes d considering only attribute sweating that is $E(a_1)_{sever/moderate}$(positive) /(negative) cases can't classified as several ambiguity result found out that is $\{E_2,E_5\}$, $\{E_9,E_{10}\}$, $\{E_{12},E_3\}$, $\{E_{14},E_{15}\}$,$\{E_{16},E_{17}\}$ with respect to decision variable d so for that sweating gives insignificant result so this attribute has hardly any importance. similarly for the symptoms of Polyuria we have to find the degree of dependency (polyuria attributes as a_2) $E(a_2)_{severe/moderate}$(positive)= $\{E_1,E_2,E_6,E_7,E_9,E_{12}, E_8, E_{15},E_{16},E_{20}\}$ so degree of dependency 10/20 for the positive cases with respect to decision variable d similarly the negative polyuria cases are $E(a_2)_{moderate}$ (positive)=$\{ E_8, E_{15},E_{16},E_{20}\}$ 4/20 $E(a_2)_{moderate//severe}$ (negative)= $\{ E_{17}, E_{18},E_{19}\}$ 3/20 for that we can generate significant result for polyuria that is whether polyuria is severe or moderate generally produces positive cases so polyuria has certain level of significance in diabetes that is if polyuria leads to diabetes now analyzing Polydipsia that is a_3 we have the following results $E(a_3)_{moderate /severe}$(positive)= $\{ E_1, E_2, E_6 E_7 ,E_8,E_9 E_{12},E_{15},E_{20}\}$ E_{16},E_{19} Produces ambiguous result so here the degree dependency 9/20 on positive cases two ambiguous cases similarly the negative cases $E(a_3)(negative)_{moderate/severe}=\{E_{10},E_{11}, E_{13},E_{14},E_{17},E_{18},E_{19}\}$

That is the degree of dependency will be 7/20 but in analyzing the data we have the cases like E_1, E_2, E_8,E_{12} produces the same result that is if Polydipsia is moderate then we have positive cases similarly analyzing the negative cases we have similar result E_5,E_6 produces ambiguous result so we are consider those and for other cases E_{10}, E_{13},E_{14},E_{17},E_{18} produces the same result that is all severe cases of Polydipsia produces negative result , that if Polydipsia is sever the person has negative cases so on analyzing the data , case of polydipsia we have insignificant result that is the person may have polydipsia then some cases gives positive result and other cases has negative result so no definite rule can be extracted from it so we drop this polydipsia attribute for further investigation. Now investigate $E(a_4)_{moderate/severe}$(positive) ={ E_1,E_6, E_7,E_{12},E_{15},E_{16}} dependency factor for positive cases will be 6/20

$E(a_4)_{normal/moderate}$ (negative)={ E_5,E_{11}, E_{14},E_{18}} E_{19},E_{20} gives ambiguous result here dependency factor for negative cases will be 4/20 similarly for analyzing Restlessness we have $E(a_5)_{severer\ /moderate}$ (positive)={E_1,E_6, E_{15}} two ambiguity result E_8, E_{13} and E_{12} ,E_{18} in negative cases similarly in negative cases E_5, E_7 are ambiguous result so need not go for further investigation so we can drop two attributes from the tables that is a_1,a_5 from the table

Leads to table-3

Table-3:

E	a_2	a_3	a_4	d
E_1	b_2	b_1	b_1	c_1
E_2	b_2	b_1	b_3	c_1
E_5	b_1	b_2	b_3	c_2
E_6	b_2	b_2	b_2	c_1
E_7	b_2	b_2	b_2	c_1
E_8	b_1	b_1	b_1	c_1
E_9	b_2	b_2	b_3	c_1
E_{10}	b_2	b_2	b_2	c_2
E_{11}	b_2	b_1	b_3	c_2
E_{12}	b_2	b_1	b_1	c_1
E_{13}	b_2	b_2	b_2	c_2
E_{14}	b_2	b_2	b_3	c_2
E_{15}	b_1	b_1	b_1	c_1
E_{16}	b_1	b_1	b_1	c_1
E_{17}	b_2	b_2	b_2	c_2
E_{18}	b_2	b_2	b_3	c_2
E_{19}	b_2	b_1	b_3	c_2
E_{20}	b_1	b_2	b_3	c_1

In table 3 we found E_1,E_{12} provides same values similarly E_6,E_7 also provide the same result and E_2,E_{11} ambiguous result so we keep one table E_1 for E_1,E_{12} and keep E_6 for E_6,E_7 and drop both E_2,E_{11} from the tables to leads to table 4

Table-4

E	a_2	a_3	a_4	d
E_1	b_2	b_1	b_1	c_1
E_5	b_1	b_2	b_3	c_2
E_6	b_2	b_2	b_2	c_1
E_8	b_1	b_1	b_1	c_1
E_9	b_2	b_2	b_3	c_1
E_{10}	b_2	b_2	b_2	c_2
E_{13}	b_2	b_2	b_2	c_2
E_{14}	b_2	b_2	b_3	c_2
E_{15}	b_1	b_1	b_1	c_1
E_{16}	b_1	b_1	b_1	c_1
E_{17}	b_2	b_2	b_2	c_2
E_{18}	b_2	b_2	b_3	c_2
E_{19}	b_2	b_1	b_3	c_2
E_{20}	b_1	b_2	b_3	c_1

From the table -4 we get conclusion that E_5,E_{20} provides ambiguous result so we drop both E_5,E_{20} from the table leads to table table-5

Table-5

E	a_2	a_3	a_4	d
E_1	b_2	b_1	b_1	c_1
E_6	b_2	b_2	b_2	c_1
E_8	b_1	b_1	b_1	c_1

E_9	b_2	b_2	b_3	c_1
E_{10}	b_2	b_2	b_2	c_2
E_{13}	b_2	b_2	b_2	c_2
E_{14}	b_2	b_2	b_3	c_2
E_{15}	b_1	b_1	b_1	c_1
E_{16}	b_1	b_1	b_1	c_1
E_{17}	b_2	b_2	b_2	c_2
E_{18}	b_2	b_2	b_3	c_2
E_{19}	b_2	b_1	b_3	c_2

Again analyzing table -5 we have E_6,E_{10} produces ambiguous result and { E_{13},E_{17} }leads to single results that is E_{13} so table -5 further reduces to table -6 by deleting the ambiguity and redundancy

Table-6

E	a_2	a_3	a_4	d
E_1	b_2	b_1	b_1	c_1
E_8	b_1	b_1	b_1	c_1
E_9	b_2	b_2	b_3	c_1
E_{13}	b_2	b_2	b_2	c_2
E_{14}	b_2	b_2	b_3	c_2
E_{15}	b_1	b_1	b_1	c_1
E_{16}	b_1	b_1	b_1	c_1
E_{18}	b_2	b_2	b_3	c_2
E_{19}	b_2	b_1	b_3	c_2

Now further classification E_{15},E_{16} leads to same class that is{ E_{15},E_{16} }= E_{15} further reduction produces table-7 by deleting the redundant rows.

Table-7

E	a_2	a_3	a_4	d
E_1	b_2	b_1	b_1	c_1
E_8	b_1	b_1	b_1	c_1
E_9	b_2	b_2	b_3	c_1
E_{13}	b_2	b_2	b_2	c_2
E_{14}	b_2	b_2	b_3	c_2
E_{15}	b_1	b_1	b_1	c_1
E_{18}	b_2	b_2	b_3	c_2
E_{19}	b_2	b_1	b_3	c_2

Continuing the reduction process we further reduces E_{14},E_{18} giving the same conclusion both leads to same result which generate the reduction table as table-8

Table-8

E	a_2	a_3	a_4	d
E_1	b_2	b_1	b_1	c_1
E_8	b_1	b_1	b_1	c_1
E_9	b_2	b_2	b_3	c_1
E_{13}	b_2	b_2	b_2	c_2
E_{14}	b_2	b_2	b_3	c_2
E_{15}	b_1	b_1	b_1	c_1
E_{19}	b_2	b_1	b_3	c_2

The same procedure again gives us further reduction that is E_8, E_{15} also leads to same information sets so futher reduction gives another tale named as table-9

Table-9

E	a_2	a_3	a_4	d
E_1	b_2	b_1	b_1	c_1
E_8	b_1	b_1	b_1	c_1
E_9	b_2	b_2	b_3	c_1
E_{13}	b_2	b_2	b_2	c_2
E_{14}	b_2	b_2	b_3	c_2
E_{19}	b_2	b_1	b_3	c_2

Here in table -9 again we have E_9,E_{14} leads to ambiguous results so dropping both the table for further classification we have table-10

Table-10

E	a_2	a_3	a_4	d
E_1	b_2	b_1	b_1	c_1
E_8	b_1	b_1	b_1	c_1
E_9	b_2	b_2	b_3	c_1
E_{13}	b_2	b_2	b_2	c_2
E_{19}	b_2	b_1	b_3	c_2

Now next we find the the strength of rules for attributes a_2, a_3, a_4 strength of rules for attributes define as strength for an association rule x→D define as is the the number of examples that contain xUD to the number examples that contains x

$(a_2=b_2)\rightarrow(d=c_1)=2/3=66\%$,$(a_2=b_1)\rightarrow(d=c_1)=1=100\%$,,$(a_2=b_2)\rightarrow(d=c_2)=2/4=25\%$, $(a_2=b_1)\rightarrow(d=c_2)$=nil now we calculate strength for a_3 $(a_3=b_1)\rightarrow(d=c_1)=2/3=66\%$,$(a_3=b_2)\rightarrow(d=c_1)=1/2=50\%$,$(a_3=b_1)\rightarrow(d=c_2)=1/3=33\%$,$(a_3=b_2)\rightarrow(d=c_2)=1/2=50$

Similarly strength for a_4 will be $(a_4=b_1)\rightarrow(d=c_1)=1=100\%$
$(a_4=b_2)\rightarrow(d=c_1)=1=100\%$,$(a_4=b_1)\rightarrow(d=c_2)$=nil
$(a_4=b_3)\rightarrow(d=c_2)=1/2=50\%$, $(a_4=b_2)\rightarrow(d=c_2)=100\%$

In this analysis we find a_2 and a_3 must important attributes in analyzing the data analysis as because we are having a result for a_4 in severe case of Polyphagia we resulted a negative result so this attributes also not important like a_2, a_3 from the above analysis we develop a rule that is

If $(\text{Polyuria})_{\text{moderate/severe}}\rightarrow$ symptoms for diabetes that is a_2 moderate or severe leads to diabetes similarly

For $(\text{polydipsia})_{\text{moderate/severe}}\rightarrow$ may leads to be diabetes because of 50% chances of negative also exit in severe plydipsia cases

Statistical validation- We basically focus on sample size for our paper , we consider a sample size of 1000 , although we get a conclusion . As rough set deals with uncertainty may leads to some kind of confusion regarding the result to validate our claims we depends upon chi squared test to validate our claim by using chi squared test

We found that chi squared value that is chi squared value we consider as k which lies below the critical range .

Future work- Our work can be extended to different fields like student feedback system , Business data analysis, Medical data analysis

Conclusion-The diagnosis process is in medical science based upon , different medical test our effort in this paper is to go by symptom of the diseases , data collected from various sources, applying rough set concept we develop the algorithm which is précised ,lucid and can be develop further .

REFERENCES

1. Cios, K., W. Pedrycz and R. Swiniarski (1998). *Data Mining Methods for Knowledge Discovery*. Kluwer Academic

2. Wolf, S., H. Oliver, S. Herbert and M. Michael (2000). Intelligent data mining for medical quality management

3. Se-Ho, Ch., and P. Rockett (2002). The training of neural classifiers with condensed datasets. *SMCB*, **32**(2), 202–206.,

4 Setiono, R. (2000). Generating concise and accurate classification rules for breast cancer diagnosis. *Artificial Intelligence in Medicine*, **18**(3), 205–219

5 Cheeseman, P., and J. Stutz (1996). Bayesian classification (AutoClass): theory and results. In U.M. Fayyad

6. Grzymala–Busse, J., Z. Pawlak, R. Slowinski and W. Ziarko (1999). Rough sets. *Communications of the ACM*

7. Hassanien, A.E. (2003). Classification and feature selection of breast cancer data based on decision tree algorithm

8 Parido, A., and P. Bonelli (1993). A new approach to fuzzy classifier systems. In *Proceedings of the FifthInternational Conference on Genetic Algorithms*. pp. 223–230

9 Lin, T.Y., and N. Cercone (1997). *Rough Sets and Data Mining*. Kluwer Academic Publishers.Ning, S., H. Xiaohua, W. Ziarko and N. Cercone (1994). A generalized rough sets model. In *Proceedings ofthe 3rd Pacific Rim International Conference on Artificial Intelligence*, Vol. 431. Beijing, China. Int. Acad.Publishers. pp. 437–443.

10.. Pawlak, Z. (1991). *Rough Sets-Theoretical Aspect of Reasoning about Data*. Kluwer Academic Publishers. Pawlak, Z., J. Grzymala–Busse, R. Slowinski, W. Ziarko (1995). Rough sets. *Communications of the ACM*