Rough Set Approach for Generation of Classification Rules for Malaria

Sujogya Mishra
Research scholar, Utkal University,
Bhubaneswar-751004, India

Shakti Prasad Mohanty
Department of Mathematics,
College of Engineering and Technology
Bhubaneswar-751003, India

Sateesh Kumar Pradhan
Department of Computer Science, Utkal University,
Bhubaneswar-751003, India

Rasheswari B Ray
Research scholar, Utkal University, Bhubaneswar-04

ABSTRACT
In recent age medical science has improved to a significant height but lots of common known dieses needs maximum number of medical test that are not only expensive and but also inaccurate in case properly diagnosis of dieses in our paper we emphasized more on symptom rather than medical test .From the large domain of dieses we consider malaria fever for our purpose .Every year millions of people died from malaria due for proper diagnosis .We develop an algorithm using rough set concept. We classified the entire paper in to three basic section 1st section about literature review the 2nd section about the experiment on the data that are collected from different medical sources and the 3rd section is about the validation, conclusion and future work . The data we consider coming out to be identical with our previous paper and the process of simulation is also almost the same

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

1. INTRODUCTION
The growth of the size of data and number of existing databases far exceeds the ability of humans to analyze this data, 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. This paper organized in the following manner 1st section contains the literature review like definition such as rough sets and elementary concepts of rough set theory, correlation and some statistical validation techniques the 2ndsection consists of data analysis of the medical data which we collected from DR P K Mishra M D and 3rd section contains the algorithm for rule generation and the classification of malaria fever and statistical validation of our result . The 4th section is about the conclusion part and future work.

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 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 $\overline{RX}$=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)= $\left|\frac{POS_C(D)}{|u|}\right|$ where POSC(D) =U 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 POSC (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 POSR(D) = POSC(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 number 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

2.8 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.EI=npi where n is the number samples considered pi is the probability of ith occurrence .Data selected at random there are two hypothesis null hypothesis and alternate hypothesis null hypothesis denoted by H0 alternate hypothesis denoted by H1. H0 is the claim does follow the hypothesis and H1 is the claim does not follow the hypothesis here H1 is called the alternate hypothesis to H0.If the test value found out to be K then K can be calculated by the formula K=∑(OI-EI)2/ EI. Choice of significance level always satisfies type 1 error .

2.9 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 system . We initially consider 1000 samples, of malaria cases and five conditional attributes such as by considering five such as Fever, Headache, Chills, Fatigue and Nausea and two decision attribute positive and negative . The data we collected about malaria positive and negative cases from Doctor Pradeep Kumar Mishra MD

4. DATA REDUCTION

As the volume of data is increases with time it is difficult to know which attributes are responsible for a particular application. The basic objective of data reduction is to find the relevant attributes that have all essential information of the data set. The process is illustrated by applying rough set concept of 20 samples which we collected by correlation techniques. In this particular problem we consider five conditional attributes as Headache, Fever, Chills, Fatigue and Nausea these and it's values are

defined low ,moderate and high decision attributes are positive , negative . For better understanding and good clarity of the paper we rename the conditional attributes and it's values as (a1,a2,a3,a4,a5),(b1,b2,b3) and the decision attributes are as (c1,c2) respectively. . Application and analysis on the data set and rule generation being presented in the following tables . Table -1 is the initial table , and the process of analysis is present in the subsequent tables

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_2
E_2	b_2	b_2	b_1	b_3	b_3	c_2
E_3	b_1	b_2	b_2	b_3	b_3	c_1
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_2
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_1
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_1
E_{15}	b_2	b_1	b_1	b_1	b_1	c_2
E_{16}	b_1	b_1	b_1	b_1	b_1	c_2
E_{17}	b_1	b_3	b_2	b_2	b_3	c_1
E_{18}	b_1	b_2	b_2	b_3	b_2	c_2
E_{19}	b_1	b_3	b_1	b_3	b_3	c_2
E20	b_1	b_3	b_1	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 E19,E20, ambiguous in nature and E3, E4 gives same result so both E19,E20 drop from the table and E3 , E4 gives the same result so E3 , E4 records we only keep only one record in the table-2.From table -1 we get the next table-2

Reduced Table-2 from 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_2
E_2	b_2	b_2	b_1	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_1
E_6	b_1	b_2	b_2	b_2	b_2	c_2
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_1
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_1
E_{15}	b_2	b_1	b_1	b_1	b_1	c_2
E_{16}	b_1	b_1	b_1	b_1	b_1	c_2
E_{17}	b_1	b_3	b_2	b_2	b_3	c_1
E_{18}	b_1	b_2	b_2	b_3	b_2	c_2

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 RL.

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 RU

Boundary Region (BR) :
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 = RU-RL.

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 positive case and another group consists of negative cases

Epositive ={E4, E5, E7, E9, E11, E12, E14 } …….(1)

Enegative ={E1, E2, E6, E8, E10, E13, E15,E16,E18 }……(2)

E(a1)low={ E4, E6, E8, E9, E10, E12, E16,E17, E18}……..(3)

E(a1)moderate={ E1, E2, E7, E11, E15 }……..(4)

E(a1)high ={ E5, E13, E14 }……..(5)

The above result when compared with the positive cases E(a1)high strength[11]

Found to be nil where as for negative cases of high E(a1) strength[11] 3/3 cent percent similarly for negative cases of moderate E(a1) strength[11] gives rise to be 1/5 about 20% , positive cases of low E(a1) strength[11] 6/9 about 66% basic observation gives rise is that a1 attribute does not give any significant result similarly

E(a2)low={ E8, E15, E16 }……..(6)

E(a2)moderate={ E1, E2, E4,E6,E7,E9,E10,E12,E13,E18 }……..(7)

E(a2)high ={ E5,E13,E14}……..(8)

Similar analysis strength[11] positive high a2 will be 4/4=1 about cent percent

And for strength[11] negative for low a2 will be also 3/3=1 about cent percent

That is we have a conclusion that if a person having high a2 symptom shows the sign of malaria and in contrast a low symptom of a2 shows a negative case

That is why we are not analyzing the moderate a2 cases now similarly for a3

E(a3)low={E1,E2,E8,E15,E16}……(9)

E(a3)moderate={E4,E6,E7,E9,E10,E13,E17,E18}……(10)

E(a3)high={E5,E11,E12,E14 }……..(11)

Finding the strength [11] high a3 will be 4/4=1 that is about cent percent

And similarly for strength [11] for low a3 cases will be gives negative result will be of 4/5 about 80% negative cases for low a3 cases so we are not consider the moderate a3 case now similarly for a4 cases we consider the

E(a4)low={E1,E8,E12,E15,E16 }……..(12)

E(a4)moderate={E6,E7,E10,E13,E17 }……..(13)

E(a4)high={E2,E4,E5,E9,E11,E14,E18 }……..(14)

Analyzing a4 E(a4)low negative strength[11] will be 4/5 that is about 80% similarly for E(a4)high positive strength[11] cases will be about 5/7 about 70% now considering a5

E(a5)low={E1,E8,E13,E15,E16 }……..(15)

E(a5)moderate={E5,E6,E7,E10,E12,E18 }……..(16)

E(a5)high= {E2,E4 ,E9, E11, E14, E17 }……..(17)

E(a5)low strength[11] for negative case will be 4/5 about 80% and E(a5)high strength for positive 4/6 about 66% positive strength[11] moderate case given by 3/8 about 37% so after analyzing the above data by strength view point we ignore attribute a1 and a5 as in a1 strength [11] for low negative case is 100% high positive strength[11] a1 about 20% similar argument in case of a5 it's strength[11] among a2, a3, a4 it is the minimum one secondly the upper approximation positive case of a5 in both moderate and high case will be {E5,E7,E12, E4 ,E9, E11, E14, E17}

And the positive lower approximation the definite positive cases will be

Expositive ={E4, E5, E7, E9, E11, E12, E14 } so the boundary region for positive case will be only a single record E17 and there are lots of ambiguity that is in high a5 there also of negative cases that is E2 similarly for the moderate cases the negative cases also lies that is E6, E10, E18 that is some cases of a5 we have also negative cases so in table-3 we drop a1,a5 so after dropping a1and a5 from table 2 we have the new reduct table, named as table-3

Reduced Table-3 from Table-2

E	a2	a3	a4	D
E1	b2	b1	b1	c2
E2	b2	b1	b3	c2
E4	b2	b2	b3	c1
E5	b3	b3	b3	c1
E6	b2	b2	b2	c2
E7	b2	b2	b2	c1
E8	b1	b1	b1	c2
E9	b2	b2	b3	c1
E10	b2	b2	b2	c2
E11	b3	b3	b3	c1
E12	b2	b3	b1	c1
E13	b2	b2	b2	c2
E14	b3	b3	b3	c1
E15	b1	b1	b1	c2
E16	b1	b1	b1	c2
E17	b3	b2	b2	c1
E18	b2	b2	b3	c2

Upon analyzing table-3 we have the following result that is (E5,E11 ,E14),

(E8,E15 ,E16), (E10,E13), forms group and (E6,E7) (E9,E9,E18),ambiguous so we keep on record for each group and delete all records which give ambiguous result so we have the new table appears as table-4 given as follows Reduced Table-4 from Table-3

Table-4 final table

E	a2	a3	a4	D
E1	b2	b1	b1	c2
E2	b2	b1	b3	c2
E5	b3	b3	b3	c1
E8	b1	b1	b1	c2
E10	b2	b2	b2	c2
E12	b2	b3	b1	c1
E17	b3	b2	b2	c1

Further reduction of Table -4 is not possible From the table we are develop an algorithm is as follows

1. High fever and chills leads to positive case of malaria.

2. Moderate fever , high chills and low fatigue leads to positive case of malaria.

3. Moderate fever, chills and fatigue leads to negative case of malaria.

4. Low fever low chills and low fatigue leads to negative case of malaria.

5.High fever moderate chills and moderate fatigue leads to positive case of malaria.

This give the idea that fever, chills and fatigue in general symptom for malaria.

Time complexly analysis- For finding the reduct we are comparing each record with another suppose there are n records then time complexity will be n+(n -1)+(n-2)+(n-3)+…….+1.=n(n-1)/2 that is of O (n2) that is the worst case analysis average case analysis of breaking the table O (nlgn) we are taking the average case analysis for dividing the table because on an average single table break down to half of it's size every time so the entire complexity will be O (nlgn+n2)

Statistical validation- For validate our findings we basically depends upon chi-square test for this purpose we consider we take a survey by taking data regarding the positive case and we are not focused on one medical centre to collect data we approached several hospital and the apply chi square test to validate our claim. . Chi square test- Expected15%,10%,15%,20%,30%,15% and the Observed samples are 25,14,34 45,62,20 so totaling these we have total of 200 samples so expected numbers of samples per each day as follows 30,20,30,40,60,30 . We then apply chi square distribution to verify our result assuming that H0 is our hypothesis that is correct H1 as alternate hypothesis that is not correct , Then we expect sample in six cases as chi squared estimation formula is $\sum$(Oi-Ei)2/ Ei where i=0,1,2,3,4,5 so the calculated as follows

X2=(25-30)2/20+(14-20)2/20+(34-30)2/30+(45-40)2/40+(62-60)2/60+(20-30)2/30
X2=25/20+36/20+16/30+25/40+4/60+100/30
=7.60
the tabular values we have with degree of freedom 5 we get result 11.04

The calculated X2 is less than the tabular value so we accept the hypothesis

5. Our work can be extended to field of agriculture , general feedback system and also to the field of entertainment.

5. 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.

[6] Fayyad 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

[11] Renu Vashist Prof M.L Garg Rule Generation based on Reduct and Core :A rough set approach International Journal of Computer Application(0975-887) Vol 29 September -2011 Page 1-4