Applied Mathematics and Computing
Vol. 1, No.1, (June, 2023)

Mathematical Modelling to extract feature for Cholera

Subrata Kumar Nayak

Dept. of Computer Science,
GVHSS, Sishu Ananta Mahavidyalaya,
Balipatna, Odisha, 751029, India
subratanayaksap@gmail.com

Sateesh Kumar Pradhan

Department of Computer Science
and Application, Utkal University,
Bhubaneswar, Odisha
sateesh1960@gmail.com

Sujogya Mishra

Dept. of Mathematics
Odisha University of Technology and Research,
Bhubaneswar,Odisha, 751029, India
sujogya123@gmail.com

Abstract: Cholera is a traditional disease with typical symptoms, In this paper our main intention to extract the exact feature for cholera by using Rough Set Theory (RST) one of the soft computing techniques. For this work, we collected nearly 10,000 samples from different parts of our state (Odisha), and using correlation techniques the size reduces to 6 number of dissimilar records. We apply Rough Set techniques on these 6-records to extract the exact feature for Cholera.

Keywords: Rough Set, Cholera, Soft computing, Feature Selection, Correlation.

1 Introduction

From the 1970s to date several People died due to Cholera [1]. Cholera symptoms are variable from the viewpoint of the Medical dictionary. Our intention to extract the exact feature for cholera that will be beneficial for the doctor in the diagnosis and treatment of cholera. To get a concrete result we use the correlation technique [2] to find the number of dissimilar records and then apply the rough set [3] technique for feature extraction. Several researchers working in this field from time to time in feature extraction /attribute reduction. Yuen Su [4] uses fish swarm and Rough set algorithm for attribute reduction, Renu Vashist[5] et al works on how to find core and reduct of Rough Set Renu Vashist and M.L Garg[6] works on Rule generation based on Reduct and Core of Rough Set. Tian-rui[7] et al study on reduct and core computation in an incompatible information system. We had collected data related to cholera from various parts our states of Odisha's interior

parts given in the form of a table after collecting the data we apply a correlation technique to reduce 10,000 records to 6-number of dissimilar records. Our major analysis is on those 6-records to extract features for cholera. We use the Rough set technique to reduce the number of conditional attributes taken initially and get a concrete set of the attribute (conditional) responsible for cholera.

1.2 Basics of Rough Set (RST)

Rough Set [8] was developed and extended from traditional Set Theory by Polish Mathematician Z.Pawlak in the year 1982. Rough deals with Vague/ Imprecise data.

Rough Set define by two basic concepts Upper Approximation, Lower Approximation.

1.3 Decision Table

Information Table-1

E	C	D
<a,b,c,d,e>	<c_1,c_2,.....c_n>	<d_1,d_2,.....>

E is the Set of records let say <a,b,c,d,e>,C is the conditional attributes and it's values are <c_1,c_2,c_3,.....> D is the decision attributes and it's values are <d_1,d_2,d_3,.....>.

Upper Approximation-: When the object of interest probably belongs to the set X. Mathematically define by the formula given below

$$\overline{B} = \{ \{x \mid [x]_B \cap X \neq \varphi\}$$

Lower Approximation-: When the object of interest belongs to the target set called as lower approximation mathematically define as

$$\underline{B}(X) = \{x : [x]_B \subseteq X\}$$

Boundary Region-: Boundary Region defines as the difference between Upper and Lower Approximation

1.3 Analytical Phase-1

Input Information Table

Place	Age	Symptoms	Total Number
Kalahandi	<15	<Watery Diarrhea, Vomiting, Thirst, Leg cramps>	5000
	15<Age<25	do	7000
	25<Age<45	do	4000
	Age>45	do	1500
Koraput	<15	do	1500
	15<Age<25	do	5000
	25<Age<45	do	3000
	Age>45	do	900
Bhabanipatana	<15	do	500

	15<Age<25	do	1500
	25<Age<45	do	2500
	Age>45	do	2500
Kendujhar	<15	do	5000
	15<Age<25	do	1500
	25<Age<45	do	1500
	Age>45	do	2500

1.4 Analytical Phase -2

In this phase, we rename the conditional attributes and decision attributes and it's valued respectively for better understanding and analysis. The renaming forms are given in the form of a Table.

watery diarrhea as p, vomiting as q, thirst as r, leg cramps s, restlessness as t, body ache as u and its values are significant and insignificant as 1 and 2, decision attributes d and its values are notable and pointless renamed as a and b respectively. For better clarity and understanding we represent all these in the form of a table. We get six records out of 10,000 records using the correlation technique. The standard records used for comparison has 6-conditional define above.

Table for Renaming of the conditional attributes and decision attributes

Conditional Attribute	Re Name of Conditional Attributes	Values of Conditional attributes	Renaming of Values of Conditional attributes	Decision Attributes d	Renaming the Values of conditional attributes
watery diarrhea	p	Significant	1	Note able	a
vomiting	q	Insignificant	2	Pointless	b
thirst	r				
leg cramps	s				
restlessness	t				
body ache	u				

1.5 Analysis Phase-3

We get 6 dissimilar records after using the correlation technique define in the form of a decision table given below.

Decision Table

E	p	q	r	s	t	u	d
E_1	1	1	1	1	1	1	a
E_2	2	2	2	2	2	2	b
E_3	1	2	1	2	1	2	b
E_4	1	1	1	2	2	2	a
E_5	1	1	2	2	1	1	a
E_6	2	2	1	2	1	1	b

Indiscernibility-:

The conditional attributes are said to be dispensable if IND(B)=IND({B-a}), otherwise Indispensable. Indiscernibility relation denoted as IND (conditional attribute).

IND(p)={{E_1,E_3,E_4,E_5},{E_2,E_6}},

IND(q)= {{E_1,E_4,E_5},{E_2,E_3,E_6}}

IND(r)= {{E_1,E_3, E_4,E_6},{E_2,E_5}}

IND(s)= {{E_1}{E_3, E_4,E_6,E_2,E_5}}

IND(t)={{E_1,E_3, E_5,E_6},{E_2,E_4}}

IND(u)={{E_1,E_5,E_6},{E_2,E_3,E_4}}

IND(p,q)={{E_1,E_4,E_5},{E_2,E_6},{E_3}}

IND(p,r)={{E_1,E_3,E_4},{E_2},{E_5},{E_6}}

IND(p,s)={{E_1},{E_3,E_4,E_5},{E_2,E_6}}

IND(p,t)={{E_1,E_3,E_5},{E_2},{E_4},{E_6}}

IND(p,u)={{E_1,E_5},{E_2},{E_3,E_4},{E_6}}

IND(q,r)={{E_1,E_4},{E_2},{E_3,E_6},{E_5}}

IND(q,s)={{E_1},{E_2,E_3,E_6},{E_4,E_5}}

IND(q,t)={{E_1,E_5},{E_2},{E_3,E_6},{E_4}}

IND(q,u)= {{E_1,E_5},{E_2,E_3},{E_4},{E_6}}

IND(r,s)= {{E_1},{E_2,E_5 },{E_3,E_4,E_6}}

IND(r,t)= {{E_1,E_3,E_6},{E_2},{E_4}{E_5}}

IND(r,u)={{E_1,E_6},{E_2},{E_3,E_4},{E_5}}

IND(s,t)= {{E_1},{E4,E2},{E_3,E_5,E_6}}

IND(s,u)= {{E_1},{E4,E_2,E_3},{E_5,E_6}}

IND(t,u)= {{E_1,E_5},{E_4,E_2},{E_3}{E_5,E_6}}

IND(p,q,r)={{E_1,E_4},{E_2},{E_3},{E_5},{E_6}}

IND(p,q,s)={{E_1},{E_2,E_6},{E_4,E_5},{E_3}}

IND(p,q,t)={{E_1,E_5},{E_2},{E_4},{E_6}, {E_3}}

IND(p,q,u)={{E_1,E_5},{E_2},{E_4},{E_6}, {E_3}}

IND(q,r,s)={{E_1},{E_2},{E_3,E_6},{E_4},{E_5}}

IND(q,r,t)={{E_1},{E_2},{E_3,E_6},{E_4},{E_5}}

IND(q,r,u)={{E_1},{E_2},{E_3},{E_4},{E_5},{E_6}}

IND(r,s,t)= {{E_1},{E_2},{E_3,E_6},{E_4},{E_5}}

IND(r,s,u)= {{E_1},{E_2},{E_3,E_4},{E_5},{E_6}}

IND(s,t,u)={{E_1},{E_2,E_4},{E_3},{E_5,E_6}}

IND(p,q,r,s)={{E_1},{E_2},{E_3},{E_4},{E_5},{E_6}}

IND(p,q,r,t)={{E_1},{E_2},{E_3},{E_4},{E_5},{E_6}}

IND(p,q,r,u)={{E_1},{E_2},{E_3},{E_4},{E_5},{E_6}}

IND(q,r,s,t)={{E_1},{E_3,E_6},{E_2},{E_4},{E_5}}

IND(q,r,s,u)={{E_1},{E_2}{E_3},{E_4},{E_5},{E_6}}

IND(r,s,t,u)= {{E_1},{E_2}{E_3},{E_4},{E_5},{E_6}}

We will have the following reduct set given below

1(r,s,t,u) 2. (p,q,r,s) 3. (p,q,r,t) 4. (p,q,r,u) 5. (q,r,s,u) 6. (r,s,t,u)

2.1 The reduct set description

Reduct Set Table

Reduct Serial	
1	(r,s,t,u)
2	(p,q,r,s)
3	(p,q,r,t)
4	(p,q,r,u)
5	(q,r,s,u)
6	. (r,s,t,u)

$$\text{Core} = \bigcap \text{Reduct} = \bigcap \{\{p,q,r,s\},\{p,q,r,t\},\{p,q,r,u\},\{q,r,s,u\},\{r,s,t,u\}\} = r$$

The attribute r present in all the reduct, so r is the core of the given conditional attributes. Our main goal to find how these reduct sets are related to the decision attributed.

Information of Reduct-1

E	r	s	t	u	d
E_1	1	1	1	1	a
E_2	2	2	2	2	b
E_3	1	2	1	2	b
E_4	1	2	2	2	a
E_5	2	2	1	1	a
E_6	1	2	1	1	b

Finding strength or confidence of conditional attributes concerning decision attribute defines as the ratio (x+d) to that of x where + stands for Union

Now calculating the strength

1. $r \to 1, d \to a$,50% $r \to 2, d \to a$ also 50%

$r \to 1, d \to b$ is 50%, $r \to 2, d \to b$ 50%

2. $s \to 1$ $d \to a$ 100%, $\to 1 \to b$ will be φ

   $s \to 2$ $d \to a$ is 40%,

3. $t \to 1, d \to a, 50\%, t \to 1, d \to b$ 50%
   $t \to 2, d \to a, 50\%, t \to 2, d \to b$ 50%

4. $u \to 1, d \to a, 50\%, u \to 1, d \to b$ 33%
   $u \to 2, d \to a, 33\%, u \to 2, d \to b$ 66%

If we consider 50% strength as bottleneck then attribute (r,s,t) conditional attributes contribution is better in comparing with u.

So the final Table is given below as follows:

Reduction Table

E	r	s	t	d
E_1	1	1	1	a
E_2	2	2	2	b
E_3	1	2	1	b
E_4	1	2	2	a
E_5	2	2	1	a
E_6	1	2	1	b

In the above Table Record, E_3 and Record E_6 are the same values so we merge the two tables into one table.

Reduction Table (Modified)

E	r	s	t	d
E_1	1	1	1	a
E_2	2	2	2	b
E_4	1	2	2	a
E_5	2	2	1	a
E_6	1	2	1	b

We find (r,s,t) are three essential attributes using AO* [9]Algorithm i.e. we just Proceed to get a single directional Optimal Solution.

Set of Rules generated by using the Final Table will be

1. $r \to 1,$ $s \to 1$,$t \to 1$ decision is a
2. $r \to 2, s \to 2, t \to 2$ decision is b
3. $r \to 1, s \to 2, t \to 2$ decision is a
4. $r \to 1, s \to 2, t \to 2$ decision is a
5. $r \to 2, s \to 2, t \to 1$ decision is a

6. $r \to 1, s \to 2, t \to 1$ decision is b

Statistical Validation

We can verify our result i.e. r,s,t are important conditional attributes in rule derivation by one dimensional χ^2 [9] method.

Null Hypothesis H_0- The three attributes r,s,t nonessential attributes in the detection of Cholera

Alternate Hypothesis H_1-: The three above attributes are essential attributes for the detection of cholera.

We consider the following data taken from different medical Sources

Outcome	E	O	$\frac{(E-O)^2}{E}$
1	6	8	2/3
2	6	5	1/6
3	6	9	3/2
4	6	2	8/3
5	6	7	1/6
6	6	5	1/6
7	6	5	1/6
8	6	4	1/3
9	6	7	1/6
10	6	8	1/3

Expected Values for the Observed frequency given by,

Expected value = $\frac{\sum x_i}{n}, i = 0 \text{ to } 10, n = 10$

Chi-Square $\chi^2(9,0.95) = 3.325, \chi^2_{0.95} = 6.33$.

So we reject the Null Hypothesis and accept the alternate hypothesis

Conclusion- Our approach is to find the extra symptoms along with the common which in general neglect. We use indiscernibility relation and Reduct of Rough Set to find essential attribute responsible for cholera and validate our claim using chi-square one-dimensional test

Future work-: Our Work can be extended to the field of Business, Entertainment, Sports, Stock Market, Time Management, and many other fields.

References

1 Weekly Epidemiological Record, 8 September 2017, vol. 92, 36 (pp. 521–536)

2 Jay L Devore Probability and Statistics for Engineering and the Sciences Cengage Publication 2012.

3 Salvatore Greco, Z. Pawlak, Roman Slowinski Bayesian Confirmation Measure within Rough Set Approach 4th International Conference, RSCTC P-(264-273)

4 Baofeng Shi, Bin Meng, Hufeng Yang, Jing Wang, Wenli Shi, A Novel Approach for Reducing Attributes and Its Application to Small Enterprise Financing Ability Evaluation Volume 2018, Article ID 1032643, P(1-17) Hindawi Complexity Wiley.

5 Renu Vashist, M L Garg, Rule Generation based on Reduct and Core: A Rough Set Approach International Journal of Computer Science Vol-29 Number-9 2011.

6 Renu Vashist, M L Garg, An Algorithm for Finding the Reduct and Core of the Consistent Dataset, International Conference on Computational Intelligence and Communication Networks 2015

7 Tian-Rui Li, Ke-Yun Qing, Ning Yang, Yang Xu, Study on Reduct and Core Computation in Incompatible Information Systems, International Conference on Rough Sets and Current Trends in Computing P-471-476

8 Z.Pawlak Rough Sets Theoretical Aspects of Reasoning about data book Springer1991

9 Ronald.E.Walpole, Raymond H.Myers, Sharon L.Myers, Keying Ye, Probability & Statistics for Engineers & Scientists 9th Edition Pearson 2012