Parameters in finding why common people become Followers of Self Declared Godman by the use of Rough Set Theory

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-751004, India

Radhanath Hota
Department of Computer Science
Orissa University of Agriculture & Techonology
Bhubaneswar-751003, India

ABSTRACT
This paper conceived looking at present scenario of our country(INDIA),Job scarcity, financial crisis of middleclass people deliver self declared god man our intention is very clear that we don't want to hurt any body's sentiments .We just apply mathematical tools

On the data which collect from various sources

Keywords
Rough Set Theory, related data, regarding terrorism Granular computing, Data mining

1. INTRODUCTION
The growth of information and number of existing databases are far exceeds the ability of common people to analyze it. Which creates a needs to ex-tract knowledge from databases[1] data regarding blind belief is so high that it's databases has accumulated large volume of information . In the present scenario intelligent method[2] means data analysis based upon strong assumptions knowledge about dependencies, probability distributions and large number of experiments which unable to derive conclusions from incomplete knowledge, or cannot proceed further with the inconsistent pieces of information. The standard intelligent method available for data analysis are neural network[3] Bayesian classifier [4] genetic algorithms[5] decision trees [6] fuzzy set [7] . Rough set theory, Professor Z. Pawlak [8] .The theory of rough sets is a mathematical tool for deriving inference from un-certain and incomplete data base information. The rough set reduction algorithms contribute to approximate the decision classes using possibly large and simplified patterns [9]. Dempster–Shafer theory or statistical methods, rough set analysis requires no external parameters and uses only those information present in the given data [10].In this paper we discusses how rough set theory is useful in analyzing large data collected from different sources , and for generating classification rules from a set of observed samples of the common people. Using rough set reduction technique we find all reducts of the data which contains the minimal subset of attributes that are associated with a class

label for classification. This paper is organized in the following manner 1st section contains the literature review 2ndsection deals with data analysis . Which we collected from various sources and 3rd section contains the algorithm ,conclusion and statistical validation present in the 4th section.

The growth of data size and number of existing databases are far exceeds the ability of humans to analyze it. Which creates an urge to ex-tract knowledge from databases[1] data regarding terrorism is so high that it's databases has accumulated large volume of information . In the present scenario intelligent method[2] means data analysis based upon strong assumptions knowledge about dependencies, probability distributions and large number of experiments which unable to derive conclusions from incomplete knowledge, or cannot proceed further with the inconsistent pieces of information. The standard intelligent method available for data analysis are neural network[3] Bayesian classifier [4] genetic algorithms[5] decision trees [6] fuzzy set [7] . Rough set theory, Professor Z. Pawlak [8] .The theory of rough sets is a mathematical tool for deriving inference from un-certain and incomplete data base information. The rough set reduction algorithms contribute to approximate the decision classes using possibly large and simplified patterns [9]. Dempster–Shafer theory or statistical methods, rough set analysis requires no external parameters and uses only those information present in the given data [10].In this paper we discusses how rough set theory is useful in analyzing large data collected from different sources , and for generating classification rules from a set of observed samples of the common people. Using rough set reduction technique we find all reducts of the data which contains the minimal subset of attributes that are associated with a class label for classification. This paper is organized in the following manner 1st section contains the literature review 2ndsection deals with data analysis . Which we collected from various sources and 3rd section contains the algorithm and rule generation for classification and statistical validation the 4th section contains the conclusion part and the 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.

2.1.1 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.

2.1.2 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.

2.1.3 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.

2.1.4 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 $\underline{R}X$ = { e ε U | [e] ε X} and The upper approximation of X by R in S is defined as $\overline{RX} = \{e \in U / [e] \cap X \neq \phi\}$ 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 $\overline{RX} = \underline{R}X$ .A set X is rough in S if its boundary set is nonempty.

2.2 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.

2.3 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.

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

2.4 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

Basic idea for the proposed work is conceived from very general blind faith or belief of the society. We initially consider 1000 samples, of very common people and seven conditional attributes such as financial crisis , lack of education, ,to achieve goal in very short period of time , life below poverty lines ,socially neglected , minority, family background it's values are defined as average ,above average and high and decision attributes are positive , negative . For better understanding we rename the attributes as $(a_1,a_2,a_3,a_4,a_5,a_6,a_7)$ as conditional attributes ,(b_1,b_2,b_3) as values of the conditional attributes and (c_1,c_2) as decision attributes respectively We collected the data for our from different sources

4. DATA REDUCTION

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.

a. 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.

b. 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

c. 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 $X \neq \phi$ the set X "Rough" is considered. In that the boundary region is BR = R_U-R_L.

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	a_6	a_7	d
E_1	b_2	b_2	b_1	b_1	b_1	b_2	b_1	c_2
E_2	b_2	b_2	b_1	b_3	b_3	b_2	b_2	c_2
E_3	b_1	b_2	b_2	b_3	b_3	b_1	b_2	c_1
E_4	b_1	b_2	b_2	b_3	b_3	b_1	b_2	c_1
E_5	b_3	b_3	b_3	b_3	b_3	b_2	b_2	c_2
E_6	b_1	b_2	b_2	b_2	b_2	b_1	b_2	c_2
E_7	b_2	b_2	b_2	b_2	b_2	b_2	b_2	c_1
E_8	b_1	b_1	b_1	b_1	b_1	b_1	b_1	c_2
E_9	b_1	b_2	b_2	b_3	b_3	b_1	b_3	c_1
E_{10}	b_1	b_2	b_2	b_2	b_2	b_1	b_2	c_2
E_{11}	b_2	b_3	b_3	b_3	b_3	b_2	b_2	c_1
E_{12}	b_1	b_2	b_3	b_1	b_2	b_1	b_2	c_1
E_{13}	b_3	b_2	b_2	b_2	b_1	b_3	b_1	c_2
E_{14}	b_3	b_3	b_3	b_3	b_3	b_3	b_3	c_1
E_{15}	b_1	b_1	b_1	b_1	b_1	b_1	b_1	c_2
E_{16}	b_1	b_1	b_1	b_1	b_1	b_1	b_1	c_2
E_{17}	b_1	b_3	b_2	b_2	b_3	b_1	b_3	c_1
E_{18}	b_1	b_2	b_2	b_3	b_2	b_1	b_2	c_2
E_{19}	b_1	b_3	b_1	b_3	b_3	b_1	b_3	c_2
E_{20}	b_1	b_3	b_1	b_3	b_3	b_1	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_{15},E_{16} are same so keep one record either from E_{15} or E_{16} similarly E_3,E_4 has same value so we keep one record either from E_3 or E_4 in the next table . E_{19} and E_{20} ambiguous in nature so we drop both from next table , that is table-2

Reduced Table-2 from table-1

E	a_1	a_2	a_3	a_4	a_5	a_7	D
E_1	b_2	b_2	b_1	b_1	b_1	b_1	c_2
E_2	b_2	b_2	b_1	b_3	b_3	b_2	c_2
E_4	b_1	b_2	b_2	b_3	b_3	b_2	c_1

E_5	b_3	b_3	b_3	b_3	b_2	b_2	c_1
E_6	b_1	b_2	b_2	b_2	b_2	b_2	c_2
E_7	b_2	b_2	b_2	b_2	b_2	b_2	c_1
E_8	b_1	b_1	b_1	b_1	b_1	b_1	c_2
E_9	b_1	b_2	b_2	b_3	b_3	b_3	c_1
E_{10}	b_1	b_2	b_2	b_2	b_2	b_2	c_2
E_{11}	b_2	b_3	b_3	b_3	b_3	b_2	c_1
E_{12}	b_1	b_2	b_3	b_1	b_2	b_2	c_1
E_{13}	b_3	b_2	b_2	b_2	b_1	b_1	c_2
E_{14}	b_3	b_3	b_3	b_3	b_3	b_3	c_1
E_{15}	b_2	b_1	b_1	b_1	b_1	b_1	c_2
E_{17}	b_1	b_3	b_2	b_2	b_3	b_3	c_1
E_{18}	b_1	b_2	b_2	b_3	b_2	b_2	c_2

$E_{high} =\{E_4, E_5, E_7, E_9, E_{11}, E_{12}, E_{14} , E_{17}\}$ ....(1)

$E_{average} =\{E_1, E_2, E_6, E_8, E_{10}, E_{13}, E_{15}, E_{18} \}$ ....(2)

$E(a_1)_{average} =\{ E_4, E_6, E_8, E_9, E_{10}, E_{12}, E_{17}, E_{18}\}$ ....(3)

$E(a_1)aboveaverage =\{ E_1, E_2, E_7, E_{11}, E_{15} \}$ ....(4)

$E(a_1)_{high} =\{ E_5, E_{13}, E_{14} \}$ ....(5)

The above result when compared with the high cases $E(a_1)_{high}$ strength[11] positive found to be 2/3 about 66% where as for average cases of positive $E(a_1)_{laverage}$ is ½ about 50% strength[11] 1/3, so we arrive at a contradiction that we have 66% high positive cases and 50% average positive cases and almost 50% above average cases positive cases so for that a_1 doesn't provide any significance similarly

$E(a_2)_{average} (negative)=\{ E_8, E_{15} \}$ ....(6)

$E(a_2)aboveaverage(negative)=\{ E_1, E_2, E_4,E_6,E_7,E_9,E_{10},E_{12},E_{13},E_{18} \}$ .....(7)

$E(a_2)_{high}(positive) =\{ E_5,E_{11},E_{17}\}$ ....(8)

Similar analysis strength[11] high a_2 will be 4/4=1 about cent percent of positive result And for strength[11] for average a_2 negative will be also 3/3=1 about cent percent this attributes provide us a strong conclusion regarding the attribute a_2, as both cases positive high and average negative result is about cent percent . So this is an important attribute for further classification .now similarly for a_3.

$E(a_3)_{average}(negative)=\{E_1,E_2,E_8,E_{15}\}$ ....(9)

$E(a_3)_{aboveaverage}=\{E_4,E_6,E_7,E_9,E_{10},E_{13},E_{17},E_{18}\}$ ....(10)

$E(a_3)_{high}(positive)=\{E_5,E_{11},E_{12}\}$ ....(11)

Finding the strength [11] high a_3 high cases will be 4/4=1

that is about cent percent and similarly for strength [11] for average a_3 cases will be gives negative result will be of 4/5 about 80% negative cases for this reason we are considering the above average cases for classification and get a conclusion that a_3 is an important attribute for further classification now similarly for a_4 cases we consider the

$E(a_4)_{average}$(negative)= $\{E_1,E_8,E_{12},E_{15}\}$ …………………………..(12)

$E(a_4)_{aboveaverage}= \{E_6,E_7,E_{10},E_{13},E_{17} \}$…………… …..(13)

$E(a_4)_{high}= \{E_2,E_4,E_5,E_9,E_{11},E_{14},E_{18} \}$……… ……..(14)

Analyzing a_4 $E(a_4)_{laverage}$ negative strength[11] will be 4/5 that is about 80% similarly for $E(a_4)_{high}$ positive strength[11] cases will be about 5/7 about 70% now considering a_5

$E(a_5)_{average}= \{E_1,E_8,E_{13}, E_{15} \}$…………………………..(15)

$E(a_5)_{aboveaverage}= \{E_5,E_6 ,E_7, E_{10}, E_{12}, E_{18} \}$……………..(16)

$E(a_5)_{high} = \{E_2,E_3 ,E_9, E_{11}, E_{14}, E_{17} \}$………………..(17)

$E(a_5)_{average}$ strength[11] for negative case will be 4/4 about cent percent and $E(a_5)_{high}$ strength for positive 4/6 about 66% positive strength[11] above average case given by 3/8 about 37% so after analyzing the above data by strength view point we ignore attribute a_1 and a_5 as in a_1 strength [11] for low negative case is 100% high positive strength[11] a_1 about 20% similar argument in case of a_5 observing it's strength[11] now upon analyzing a_6 it has the following information

$E(a_6)_{laverage}=\{$ $E_4,$ $E_6,$ $E_8,$ $E_9,$ $E_{10},$ $E_{12},$ $E_{17},$ $E_{18}\}$…………..(18)

$E(a_6)_{aboveaverage}=\{$ $E_1, E_2, E_7, E_{11}, E_{15} \}$…………..(19)

$E(a_6)_{high} =\{ E_5, E_{13}, E_{14} \}$…………..(20)

Provide the following strength cases $E(a_6)_{high}$ strength[11] found to be 2/3 about 66% where as for negative cases of $E(a_6)_{average}$ is ½ about 50% strength[11] 1/3 cent percent similarly for negative cases of aboveaverage $E(a_6)$ strength[11] gives rise to be 3/5 about 60% , so we arrive at a contradiction that we have 66% positive and 50% negative and 60% negative case in moderate case of a_6 doesn't provide any significance similarly in case of a_7 there is peculiar case arises that above average cases of a_7 provide both positive and negative result so we arrive at an ambiguity ,that's why we don't analyze a_7 further

In table-3 we drop a_1,a_5,a_7 . So after dropping $a_1,$ $a_5,$ and a_7 from table 2 we have the new reduct table, named as table-3.

Reduced Table-3 from Table-2

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

Upon analyzing table-3 we have the following result that is (E_5,E_{11},E_{14}), (E_8,E_{15},E_{16}), (E_{10},E_{13}), forms group and (E_6,E_7) (E_9,E_9,E_{18}),ambiguous so we keep on record for each group and delete all records which gives ambiguous result so we have the new table appears as table-4 given as follows

Reduced Table-4 from Table-3

E	a_2	a_3	a_4	D
E_1	b_2	b_1	b_1	c_2
E_2	b_2	b_1	b_3	c_2
E_5	b_3	b_3	b_3	c_1
E_8	b_1	b_1	b_1	c_2
E_{10}	b_2	b_2	b_2	c_2
E_{12}	b_2	b_3	b_1	c_1
E_{17}	b_3	b_2	b_2	c_1

As we observe E_8 and E_{10} has conditional attributes values different but giving same decision so we safely drop both for further classification .further reduction Table -4 is not possible.

From the table we are develop an algorithm is as follows

1. above average cases of a_2 , a_4 and a_3 giving significant result for negative case of followers for self declared God man

2. above average cases of a_2,a_4 and a_3 provide a leads for negative followers for self declared god man

3. High or severe cases of a_2, a_4 and a_3 leads to

positive cases of followers for self declared god man

4. Average a_2, a_4 and , a_3 not provided significance result for followers of self declared god man

5. Above average a_2, a_3 is significant average a_4 leads to positive or definite case of followers of self declared god man

6. High a_2, a_3 above average cases of, a_4 leads to positive case of self declared god man

The above algorithm gives us a conclusion that 1) lack of education 2) to achieve goal in short period of time are root cause in the transformation of common people to become followers of self declared God man

Statistical validation I- To validate our findings we basically depends upon chi-square test for this purpose we take a survey by taking data regarding the positive cases and we are not focused on one particular areas to collect the data we approached several sources and gathered information and then 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 H_0 is our hypothesis that is correct H_1 as alternate hypothesis that is not correct , Then we expect sample in six cases as chi squared estimation formula is $\sum(O_i-E_i)^2/ E_i$ where i=0,1,2,3,4,5 so the calculated as follows

$X^2=(25-30)^2/30+(14-20)^2/20+(34-30)^2/30+(45-40)^2/40+(62-60)^2/60+(20-30)^2/30$

$X^2=25/20+36/20+16/30+25/40+4/60+100/30$

=7.60 the tabular values we have with degree of freedom 5 we get result 11.04. This result is well below the tabular values , this gives us to accept our claim.

Statistical validation II- Apply chi square test to validate our claim. For the second case as follows . Chi square test- Expected14%,16%,15.5%,21%,31%,14% and the samples are 26,18,34, 43,61,20. so totaling these we have total of 200 samples so expected numbers of samples per each day as follows 28,32,31,42,62,28 . We then apply chi square distribution to verify our result assuming that H_0 is our hypothesis that is correct H_1 as alternate hypothesis that is not correct , Then we expect sample in six cases as chi squared estimation formula is $\sum(O_i-E_i)^2/ E_i$ where i=0,1,2,3,4,5 so the calculated as follows 8.879 it is far below the tabular values 11.04 wit degrees of freedom 5 so we accept our claim.

5. FUTURE WORK
This idea can be extend to entertainment, Business logic development and in the Field of agriculture, Entertainment and Education.

6. 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.,

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

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

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

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

[7] 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

[8] 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.

[9] 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*

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

[11] Sujogya Mishra , shakthi Prasad mohanty,Sateesh Kumar Pradhan Radhanath Hota Parameters to find the cause of Global Terrorism using Rough Set Theory *International Journal of Computer Applications (0975 – 8887) Volume 127 – No.14, October 2015* Page 46-50

[12] Sujogya Mishra , shakthi Prasad mohanty,Sateesh Kumar Pradhan Radhanath . Importance of Location Classification using Rough Set Approach for the Development of Business Establishment *International Journal of Computer Applications (0975 – 8887) International Journal of Computer Applications (0975 – 8887) Volume 119 – No.24, June 2015*