Why there is so much difference between Contractual and Regular Employees in a Government Run Organization by using Rough Set

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

Radhanath Hota
Department of Computer Science,
College of Basic Science & Humanities, OUAT,
Bubaneswar

Anshuman Mishra
Department of Computer Science,
Krupajal Engineering College,
Bhubaneswar

ABSTRACT

In our country many college and school run by government agencies. There are two categories of employment, one is regular and the other one is contractual. The idea of this paper is conceived looking at violation of human rights in these places. Regular employees enjoy all the facilities such as library, job security air condition chambers in contrast contractual employees deprived from all these facilities. Our intention to find the parameter for why the above said happened by the use of rough set theory ..

Keywords

Rough Set Theory, data analysis , Granular computing, Data mining

1. INTRODUCTION

The basic idea conceived looking at the present position of the contractual employees of our country . Where all facilities provide by the government only avail by the regular employees of the college in the contrast contractual employees deprived from all these . Our intention is very clear the algorithm which we develop provide us why such disparities occur . In this context we observe several government agencies of course the contract employees. In the process we generate so much data which generation not only put us in dilemma but also it creates obstacle for us to derive the exact result . This has created an obvious challenge for us in the development of reduction of data set and to derive the exact data for a particular application. The application of rough set theory has a prime role to play for knowledge discovery in data base(s).The ever increasing field of knowledge discovery (KD) that helps in derivation of hidden information from large database[3]. Data mining is also considered as essential tool in this knowledge discovery process which uses techniques from different disciplines ranging from machine learning, statistics information sciences, database, visualization ([4]-[12]). Further, prediction of system failure needs a systematic and scientific study. The first approach to predict system failure (any establishment which fails due to lack of administration)started in 1995 by Zopounidis([24]-[26]). The methods proposed are the "five C" methods, the "LAPP" method, and the "credit-men" method. Then, financial ratios methodology was developed to counter failure prediction problem. This approach gives rise the methods for general failure prediction based on multivariate statistical analysis (Altman ([13]-[15]), Beaver[17], Courtis[18]). Frydman *et al*[19] first employed recursive *et al*[16], multi-factor model by Vermeulen *et al*[23] are also other methods developed to counter failure prediction.

This paper presents a methodology to generate certain basic attributes which actually responsible for this disparities , reduction of attributes using rough set theory. portioning, while Gupta *et a*l[20] use mathematical programming as an alternative to multivariate discriminate analysis for system failure prediction problem. Other methods used were survival analysis by Luoma, Laitinenl[21] which is a tool for company failure prediction, expert systems by Messier and Hansen[22] , neural network by Altman

2. PRILIMINARIES

2.1 Rough set

Rough set theory as introduced by Z. Pawlak[2] 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 the non-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

$\bar{R}X = \{e \in U / [e] \cap X \neq \phi\}$The upper approximation of X by R in S is defined as

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

$^{-}$RX= $\underline{R}$X, 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

$$k = \gamma(C, D) = \frac{\mathrm{I}POS_C(D)\mathrm{I}}{\mathrm{I}U\mathrm{I}}$$

$$where \quad POS_C(D) = \bigcup_{X \in U/D} \underline{C}(X)\, called\ a\ positive\ region\ of\ the$$

$$partition\ \ U/D\ \ with\ respect\ to\ C, which\ is\ the\ set\ of\ all\ elements\ of$$

$$\mathrm{U}\, that\ can\ be\ uniquely\ classified\ to\ blocks\ of\ the\ partition\ U/D$$

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.

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 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 hypothesis 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 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 $K=\sum(O_I-E_I)^2/E_I$. 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 born from the general work culture of our country . We at first consider 1000 samples, and using five conditional attributes such as 1.employer attitude ,2.salary difference, 3.Risk of losing the job for contractual employees 4.Service rule is not properly defined 5 Government policy and two decision attributes such as insignificant and significant. .For this purpose we initially consider 1000 samples then by using correlation techniques , only 20 samples are selected which appears to be dissimilar then by applying rough set concept we reduced the number of attributes . The values of the attributes define as high, moderate and low .

4. DATA REDUCTION

As the volume of data is increasing every day , it is very difficult to find which type of data is important to decide which are actual responsible for decision making .The aim of data reduction is to find the relevant attributes(for decision making) that have all essential information of the data set. The process is illustrated through the following 20 samples by using the rough set theory. For this paper we consider the conditional attributes that described in section 3 which can be applied to all types of Government run organization? To represent all these in the tabular form we rename the five the conditional attributes employer attitude as a_1 , salary difference as a_2 , Risk of losing the job for contractual employees as a_3 , Service rule is not properly defined as a_4 , Government policy as a_5. Conditional attribute values are consider as, high , moderate ,and low renamed as b_1, b_2 and b_3 respectively decision attribute d are considered as significant and insignificant renamed as c_1 and c_2 respectively. Data are collected from various sources.To start with we consider initial table which is generated from 20 samples which we get by the method of correlation techniques.

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 as our 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

4.1 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.

4.2 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.

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 groups consist of all positive case and other one all negative cases

$E_{(\text{Significant})}$={ E_1, $E_{2,}$ $E_{6,}$ $E_{7,}$ $E_{8,}$ E_9 ,$E_{12,}$ $E_{15,}$ $E_{16,}E_{20}$}…(1)

$E_{(\text{insignificant}}$)={ E_5, $E_{10,}$ $E_{11,}$ $E_{13,}$ E_{14} ,$E_{17,}$ $E_{18,}$ }............(2)

Here in this case lower approximation for Success represented by the first equation and lower approximation for failure represented by the Second equation now we find the entities which are falls into different groups to generate different equivalence classes as follows . Another notation in this case we follows as partially available denoted as p available, not available denoted as n available

$E(a_1)_{b1}$ ={E_6, $E_{8,}$ $E_{9,}$ $E_{10,}$ $E_{12,}$ E_{16} ,$E_{17,}$ $E_{18,}$ E_{19}}

$E(a_1)_{b2}$ ={E_1, E_2, E_7, E_{11} ,$E_{15,}$ E_{20}}

$E(a_1)_{b3}$ =={E_5, $E_{13,}$ E_{14} }, $E(a_2)_{high}$ ={E_8, $E_{15,}$ E_{16} }

$E(a_2)_{b1}$ ={ E_1, $E_{2,}$ E_6,E_7,E_9,E_{10},E_{12},$E_{13,}$$E_{18}$,$E_{20}$ }

$E(a_2)_{b2} =\{ E_1,E_{8,}E_{12,}E_{15}, E_{16} \}$

$E(a_3)_{b1} =\{ E_1,E_{8,}E_{12,}E_{15}, E_{16} \}$

$E(a_3)_{b2} =\{ E_6,E_7,E_{10}, E_{13} ,E_{17} \}$

$E(a_3)_{b3} =\{ E_{5,}E_{11},E_{12},E_{14}\}$ $E(a_4)_{available}=\{ E_{5,}E_{11},E_{12},E_{14} \}$

$E(a_4)_{b1} =\{ E_1,E_{8,}E_{12,}E_{15}, E_{16} \}$

$E(a_4)_{b3} =\{ E_6,E_7,E_{10}, E_{13} ,E_{17} \}$

$E(a_5)_{b1} =\{E_1,E_8,E_{13},E_{15},E_{16},E_{19}\}$

$E(a_5)_{b2} =\{E_5,E_6,E_7,E_{10},E_{12},E_{18}\}$

$E(a_5)_{b3} =\{E_2,E_9,E_{11},E_{14},E_{17},E_{20}\}$

Next, we find the combination of two attributes each to generate the reduct such combinations are $E(a_1,a_2)$, $E(a_1,a_3)$, $E(a_1,a_4)$, $E(a_1,a_5)$ $E(a_1,a_2)_{b1}=\{E_8,E_{16}\}$, $E(a_1,a_2)_{b2}=\{E_1,E_2,E_7,E_{20}\}$

$E(a_1,a_2)_{b3}=\{E_3,E_{14} \}$ $E(a_1,a_3)_{b1}=\{E_8,E_{16},E_{19}\}$ $E(a_1,a_3)_{b2}=\{E_7,E_{20}\}$

$E(a_1,a_3)_{b3}=\{E_5,E_{14}\}$ $E(a_{1,}a_4)_{b1}=\{E_8,E_{12},E_{16}\}$ $E(a_{1,}a_4)_{b2}=\{E_7 \}$ $E(a_{1,}a_4)_{b3}=\{E_5,E_{14} \}$ $E(a_{1,}a_5)_{b1}=\{E_8,E_{12},E_{16}\}$ $E(a_{1,}a_5)_{b2}=\{E_7\}$ $E(a_{1,}a_5)_{b3}=\{E_{14}\}$

$E(a_{2\,,}\ a_3)_{b1} =\{E_{8,}E_{15},E_{16}\}$

$E(a_2,,a_3)_{b2}=\{E_{6,}E_7,E_{9,},E_{10},E_{13},E_{18},E_{20}\}$,

$E(a_2,,a_3)_{b3}=\{E_{5,}E_{11},E_{14}\}$ $E(a_2,,a_4)_{b1} =\{E_{8,}E_{15},E_{16}\}$

$E(a_2,,a_4)_{b2}=\{E_{6,}E_7,E_{10},E_{13}\}$ $E(a_2,,a_4)_{b3} =\{E_5,E_{11},E_{14} \}$ $E(a_2,,a_5)_{b1}=\{E_8,E_{16} \}$ $E(a_2,,a_5)_{b2}=\{E_7\}$

$E(a_2,,a_5)_{b3}=\{E_{11},E_{14},E_{17}\}$

$E(a_3,,a_4)_{b1} =\{E_1,E_8,E_{15},E_{16}\}$ $E(a_3,,a_4)_{b2} =\{E_6,E_7,E_{1\backslash 0},E_{13},E_{17}\}$

$E(a_3,,a_4)_{b3} =\{E_5, E_{11},E_{14}\}$ $E(a_3,,a_5)_{b1} =\{E_1, E_8,E_{15},E_{16},E_{19}\}$

$E(a_3,,a_5)_{b2} =\{E_6, E_7,E_{10},E_{18} \}$ $E(a_3,,a_5)_{b3} =\{E_{11}, E_{14} \}$

$E(a_4,,a_5)_{b1} =\{E_1, E_8 ,E_{15},E_{16}\}$ $E(a_4,,a_5)_{b2} =\{E_6, E_7 , E_{10} \}$ $E(a_4,,a_5)_{b3} =\{E_2,E_9,E_{20}\}$

$E(a_1,a_2 ,a_3)_{b1} =\{E_8,E_{16}\}$ $E(a_1,a_2 ,a_3)_{b2} =\{E_7,E_{20}\}$

$E(a_1,a_2 ,a_3)_{b3} =\{E_5,E_{14}\}$

$E(a_2,a_3 ,a_4)_{b1} =\{E_8 ,E_{15}, E_{16}\}$

$E(a_2,a_3 ,a_4)_{b2} =\{E_6 ,E_7, E_{10} ,E_{13}\}$

$E(a_2,a_3 ,a_4)_{b3} =\{E_5 ,E_{11}, E_{14} \}$ $E(a_3,a_4 ,a_5)_{b1}=\{E_1,E_8, E_{15} \}$

$E(a_3,a_4 ,a_5)_{b2} =\{E_6,E_7, E_{10} \}$ $E(a_3,a_4 ,a_5)_{b3} =\{E_{11},E_{14}\}$

$E(a_1,a_2,a_3,a_4)_{b1} =\{E_8,E_{16}\}$ $E(a_1,a_2,a_3,a_4)_{b2} =\{E_7\}$

$E(a_1,a_2,a_3,a_4)_{b3} =\{E_5,E_{14}\}$

These equivalence classes are 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 print media that is $E(a_1)_{b1/b2}$ (significant) or(insignificant) 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 a_1 gives insignificant result so this attribute has hardly any importance. similarly for a_2 we have to find the degree of dependency

$(E(a_2)_{b1/b2}$ (significance)= $\{E_1,E_2,E_6,E_7,E_9,E_{12}, E_8, E_{15},E_{16},E_{20}\}$ so degree of dependency 10/20 for the success cases with respect to decision variable d similarly the insignificance cases in a_2 cases are $E(a_2)_{b2}$(significance)=$\{ E_8, E_{15},E_{16},E_{20}\}$ 4/20 $E(a_2)_{b1}$(insignificance)= $\{ E_{17}, E_{18},E_{19}\}$ 3/20 for that we can have a significant result for a_2 is high , moderate, low is in general has certain degree of significance in for a particular result is available leads to a positive or a significant result. Upon analyzing a_3 results $E(a_3)_{b1}/_{b2}$ (significance)

= $\{ E_1, E_2, E_6\ E_7 ,E_8,E_9\ E_{12},E_{15},E_{20}\}$ E_{16},E_{19} Produces ambiguous result so here the degree dependency 9/20 on significant cases two ambiguous cases similarly the negative cases $E(a_3)(\text{insignificant})_{b1/b2}$ =$\{E_{10},E_{11}, E_{13},E_{14},E_{17},E_{18},E_{19}\}$ 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 in a_3 cases is high still we have insignificant result then we have significant cases similarly analyzing the insignificant cases we have similar result E_5,E_6 produces ambiguous result so we are consider these and for other cases E_{10}, E_{13},E_{14},E_{17},E_{18} produces the same result so upon analyzing the data a_3 produces insignificant result that is in some cases this attribute produce significance and in some cases it deliver insignificance result the number in both cases are nearly equal .So for that in case of the a_3 does not provide any information from which we can generate any definite rule dropping this attribute from the decision table may hamper the investigation process so we keep this attribute in the decision table for further investigation next we investigate

$E(a_4)_{b2/b1}$(significance) =$\{ E_1,E_6, E_7,E_{12},E_{15},E_{16}\}$ dependency factor in this cases will be 6/20

$E(a_4)_{b2}/_{b1}$ (insignificance)=$\{ E_5,E_{11}, E_{14},E_{18}\}$ E_{19},E_{20} gives ambiguous result here dependency factor for negative cases will be 4/20 similarly upon analyzing we have $E(a_5)$b1/b2 (significance)=$\{E_1,E_6, E_{15}\}$ two ambiguity result E_8, E_{13} and E_{12} ,E_{18} in failure 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 so we are having new table given below . We are considering the definite cases whether insignificance or significance the cases where we are not sure of the result we keep those attribute in the table for further investigation, the reduct table which we generate presented in 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[27] 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 that is available gives a failure result so the conditional attribute a_4 is not that important like a_2,a_3 from the above analysis we develop a rule that is

Rule

1. (a_2) high/moderate shows values of a_2 provide significance result

2. For (a_3) high/moderate may leads to a significant result but still there is a 50% chances of insignificance result ..

5. STATISTICAL VALIDATION

We basically focus on colleges that has run by government agencies , 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

5.1 Experimental section

We take survey of different types of government agencies adopting the rule generated by rough set principle are as follows

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

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

=7.60 the tabular values we have with degree of freedom 5 we get result 11.04

Our experiment result is lies quite below the tabular values ,so it lies in the acceptable region .So we accept the hypothesis H_0 that of our experiment result is correct.

6. FUTURE WORK

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

7. CONCLUSION

This is based upon mathematical analysis and experiment which is gives some accurate result in generating rules, from a vast diversified data base. This also give a very précised result

8. REFERENCES

[1] S.K. Pal, A. Skowron, Rough Fuzzy Hybridization: A new trend in decision making, Berlin, Springer-Verlag, 1999

[2] Z. Pawlak, “Rough sets”, International Journal of Computer and Computer and Information Sciences, Vol. 11, 1982, pp.341–356

[3] Z. Pawlak, Rough Sets: Theoretical Aspects of Reasoning about Data, System Theory, Knowledge Engineering and problem Solving, Vol. 9, The Netherlands, Kluwer - Academic Publishers, Dordrecht, 1991

[4] Han, Jiawei, Kamber, Micheline, Data Mining: Concepts and Techniques. San Franciso CA, USA, Morgan Kaufmann Publishers, 2001

[5] Ramakrishnan, Naren and Grama, Y. Ananth, “DataMining: From Serendipity to Science”, IEEE Computer, 1999, pp. 34-37.

[6] Williams, J. Graham, Simoff, J. Simeon, DataMining Theory, Methodology, Techniques, andApplications (Lecture Notes in Computer Science/ LectureNotes in Artificial Intelligence), Springer, 2006.

[7] D.J. Hand, H. Mannila, P. Smyth, Principles ofData Mining. Cambridge, MA: MIT Press, 2001

[8] D.J. Hand, G.Blunt, M.G. Kelly, N.M.Adams, “Data mining for fun and profit”, Statistical Science, Vol.15, 2000, pp.111-131.

[9] C. Glymour, D. Madigan, D. Pregibon, P.Smyth, “Statistical inference and data mining”, Communications of the ACM, Vol. 39, No.11,1996, pp.35-41.

[10] T.Hastie, R.Tibshirani, J.H. Friedman, Elements of statistical learning: data mining, inference and prediction, New York: Springer Verlag, 2001

[11] H.Lee, H. Ong, "Visualization support for data Mining", IEEE Expert, Vol. 11, No. 5, 1996, pp. 69-75.

[12] H. Lu, R. Setiono, H. Liu,"Effective data Mining using neural networks", IEEE Transactions on Knowledge and Data Engineering, Vol. 8, No. 6, 1996, pp. 957-961.

[13] E.I Altman, "Financial ratios, discriminants analysis and prediction of corporate bankruptcy", The journal of finance, Vol. 23 , 1968, pp.589-609

[14] E.I.Altman, R.Avery, R.Eisenbeis, J. Stnkey, "Application of classification techniques in business, banking and finance. Contemporary studies in Economic and Financial Analysis", vol.3, Greenwich, JAI Press,1981.

[15] E.I Altman, "The success of business failureprediction models: An international surveys", Journal of Banking and Finance Vol. 8, no.2, 1984, pp.171-198

[16] E.I Altman, G. Marco, F. Varetto, "Corporate distressdiagnosis: Comparison using discriminant analysis and neural networks", Journal of Banking and Finance, Vol. 18, 1994, pp. 505-529

[17] W.H Beaver, "Financial ratios as predictors of failure. Empirical Research in accounting : Selected studies", Journal of Accounting Research Supplement to Vol 4, 1966, pp.71-111

[18] J.K Courtis, "Modelling a financial ratios categoric frameWork", Journal of Business Finance and Accounting, Vol. 5, No.4, 1978, pp71-111

[19] H.Frydman, E.I Altman ,D-lKao, "Introducing recursive partitioning for financial classification: the case of financial distress", The Journal of Finance, Vol.40, No. 1, 1985, pp. 269-291.

[20] Y.P.Gupta, R.P.Rao, P.K. , Linear Goal programming as alternative to multivariate discriminant analysis a note journal of business fiancé and accounting Vol.17, No.4, 1990, pp. 593-598

[21] M. Louma, E, K. Laitinen, "Survival analysis as a tool for company failure prediction". Omega, Vol.19, No.6, 1991, pp. 673-678

[22] W.F. Messier, J.V. Hanseen, "Including rules for expert system development: an example using default and bankruptcy data", Management Science, Vol. 34, No.12, 1988, pp.1403-1415

[23] E.M. Vermeulen, J. Spronk, N. Van der Wijst., The application of Multifactor Model in the analysis of corporate failure. In: Zopounidis,C.(Ed), Operational corporate Tools in the Management of financial Risks, Kluwer Academic Publishers, Dordrecht, 1998, pp. 59-73

[24] C. Zopounidis, A.I. Dimitras, L. Le Rudulier, A multicriteria approach for the analysis and prediction of business failure in Greece. Cahier du LAMSADE, No. 132, Universite de Paris Dauphine, 1995.

[25] C. Zopounidis, N.F. Matsatsinis, M. Doumpos, "Developing a multicriteria knowledge-based decision support system for the assessment of corporate performance and viability: The FINEVA system, "Fuzzy Economic Review, Vol. 1, No. 2, 1996, pp. 35-53.

[26] C. Zopounidis, M. Doumpos, N.F. Matsatsinis, "Application of the FINEVA multicriteria knowledge-decision support systems to the assessment of corporate failure risk", Foundations of Computing and Decision Sciences, Vol. 21, No. 4, 1996, pp. 233-251

[27] 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.