ISSN 2321 3361 © 2017 IJESC

Research Article Volume 7 Issue No.3

Rough Set on Clustering

S.Robinson Chellathurai[1], L.Jesmalar[2]
Associate professor[1], Assistant professor[2]
Scott Christian College, Autonomous, Nagercoil, India[1]
Holy Cross College, Autonomous, Nagercoil, India[2]

Abstract:
The chief objective of this study is to show the usefulness of Rough set theory in Data mining. The aim of this paper is to introduce the concept of the lower and upper approximation of a set can be used in clustering.

Key words: lower and upper approximation, Rough clustering, Rough centroid

1. INTRODUCTION

The goal of clustering is to group similar objects in one cluster and dissimilar objects in different clusters. A core concept of rough sets is the lower and upper approximation of a set can be used in clustering. Rough clusters are shown to be useful for representing groups of highway sections, web users, and supermarket customers. Clustering observed earthquake epicenters to identify dangerous zones.

2. DEFINITION AND MAIN RESULTS

2.1 Definition –Rough set

A *rough set* is a formal approximation of a crisp set in terms of a pair of sets which the lower and upper approximation of the original set. Let U denote the set of objects called universe and let R be an equivalence relation on U. Then (U, R) is called an approximation space. For $u, v \in$U & $(u, v) \in$ R, u and v belong to the same equivalence class it is denoted by U/R and we say that they are indistinguishable. The relation R is called an indiscernibility relation. Let $[x]_R$ denote an equivalence class of R containing element x, then lower & upper approximation for a subset X ⊆ U ,denoted by $\underline{R}(X)$ &$\overline{R}(X)$ respectively

Where $\underline{R}(X) = \{x \in U/[x]_R \subset X\}$
$\overline{R}(X) = \{x \in U/[x]_R \cap X \neq \phi\}$

Thus if an object $x \in \underline{R}(X)$ then "x surely belongs to X"
If $x \in \overline{R}(X)$ then "x possibly belong to X"

If $\underline{R}(X)$& $\overline{R}(X)$ are sets then
$R(X) = (\underline{R}(X), \overline{R}(X))$ is called a rough set with respect to R.

Definition 2.3.
The *rough membership* can be interpreted as a degree that x belongs to X in view of information about x expressed by R.
$\underline{R}(X) = \{x \in U: \mu_X(x) = 1\}$
$\overline{R}(X) = \{x \in U: \mu_X(x) > 0\}$

Results 2.4.

a) If $\overline{R}(X) = \underline{R}(X)$ then X is definable in U

b) If $\overline{R}(X) \neq \underline{R}(X)$ then X is undefinable (roughly definable) (or) **Rough set** in U
   i. If $\underline{R}(X) \neq \phi and \overline{R}(X) \neq U$ then X is undefinable (roughly definable)
   ii. If $\underline{R}(X)) \neq \phi and \overline{R}(X) = U$ then X is internally undefinable (R_i)
   iii. If $\underline{R}(X) = \phi and \overline{R}(X) \neq U$ then X is externally undefinable (R_e).
   iv. If $\underline{R}(X) = \phi and \overline{R}(X) = U$ then X is totally undefinable (R_t).

Example 2.5
Let U={0,1,2,…8 } and the relation R on U is a≡b(mod3) for all $a, b \in U$.

Let X⊆U, then the rough set of X is $R(X) = (\underline{R}(X), \overline{R}(X))$

Let R be an equivalence relation. aRb iff a ≡ b(mod3)

The equivalence classes are {{0,3,6 },{1, 4,7},{2,5,8}}

Let X ={0,2,3,5,8} then
$\underline{R}(X)$={x∈U/$[x]_R \subset$ X} ={2,5,8}
$\overline{R}(X)$={ x∈U/$[x]_R \cap X \neq \phi$} ={0,2,3,5,6,8}

X is undefinable (roughly definable)

The membership value of X is $\mu(X) = \frac{|\underline{R}(X)|}{|\overline{R}(X)|} = \frac{3}{6}$

The membership value of each element of X in U is $\mu_X(x) = \frac{|([x]_R \cap X)|}{|[x]_R|}$

$\mu_X(0) = \frac{|([0]_R \cap X)|}{|[0]_R|} = \frac{2}{3}$, $\mu_X(1) = \frac{|([1]_R \cap X)|}{|[1]_R|} = 0$, $\mu_X(2) = \frac{|([2]_R \cap X)|}{|[2]_R|} = 1$,

$\mu_X(3) = \frac{|([3]_R \cap X)|}{|[3]_R|} = \frac{2}{3}$, $\mu_X(4) = \frac{|([4]_R \cap X)|}{|[4]_R|} = 0$, $\mu_X(5) = \frac{|([5]_R \cap X)|}{|[5]_R|} = 1$,

$\mu_X(6) = \frac{|([6]_R \cap X)|}{|[6]_R|} = \frac{2}{3}$, $\mu_X(7) = \frac{|([7]_R \cap X)|}{|7]_R|} = 0$, $\mu_X(8) = \frac{|([8]_R \cap X)|}{|[8]_R|} = 1$

3. ROUGH CLUSTERING

The goal of clustering is to group similar objects in one cluster and dissimilar objects in different clusters. A core concept of rough sets is the lower and upper approximation of a set can be used in clustering. Rough clusters are shown to be useful for representing groups of highway sections, web users, and supermarket customers. Clustering observed earthquake epicenters to identify dangerous zones.

The following method is another simple method for rough clustering.

Let us consider $U/D = \{C_1, C_2, \ldots, C_k\}$ that partitions the set U based on an equivalence relation D, it is possible to define each set $C_i \in U/D$ using its lower $\underline{R}(C_i)$ and upper $\overline{R}(C_i)$ approximations.We are considering the upper and lower bounds of only a few subsets of U. However, the family of upper and lower bounds of $C_i \in$U/D are required to follow some of the basic rough set properties such as:

Table.1. Eagerness to learn

U	Age	Eagerness to Learn	Performance (Decision)
x_1	16	Good	Good
x_2	14	Good	Good
x_3	15	Good	Good
x_4	14	Fair	Good
x_5	15	Bad	Fair
x_6	16	Fair	Fair
x_7	14	Bad	Fair
x_8	16	Bad	Bad
x_9	16	Good	Bad
x_{10}	16	Bad	Bad

The set of attributes A = {Age, EL} is considered. We have the following equivalence classes

$\{x_1,x_9\}, \{x_2\}, \{x_3\}, \{x_4\}, \{x_5\}, \{x_6\}, \{x_8,x_{10}\}$

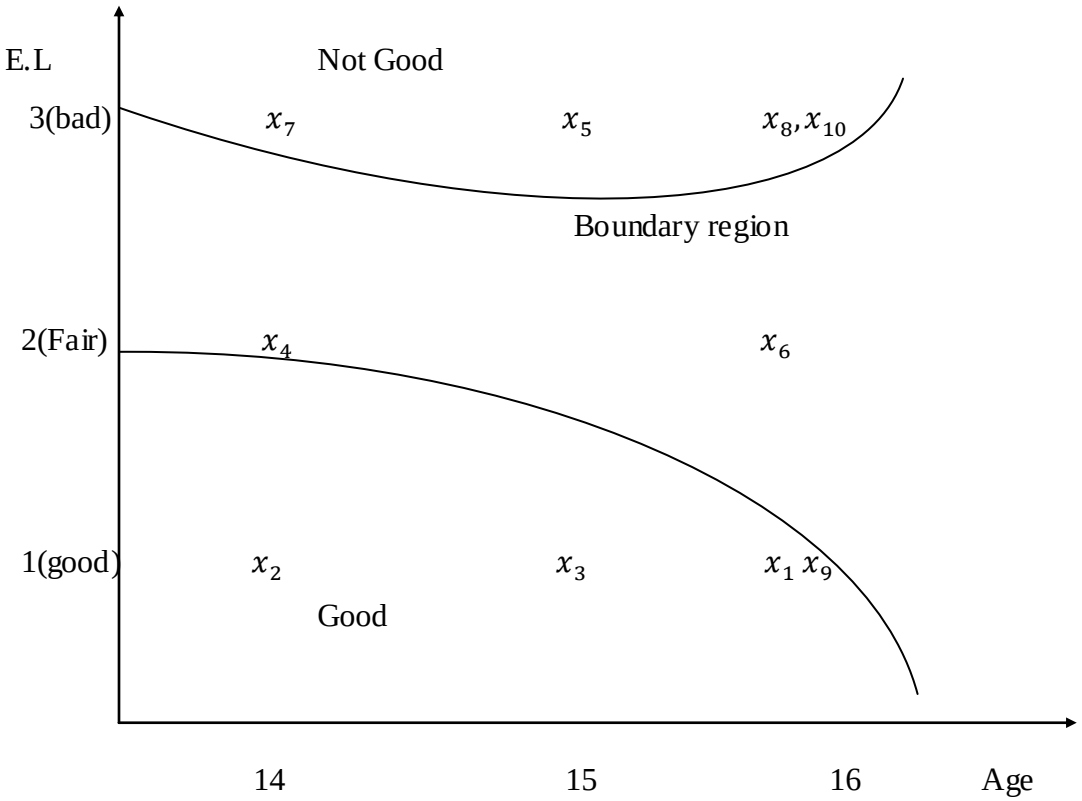

Figure.1: Rough set results

4. 1Calculation of the Centroids.

Let us assume that the objects are represented by m-dimensional vectors. The objective is to assign these n objects to k clusters. Each of the cluster is also represented by an m-dimensional vector, which is the centroid or mean vector for that cluster. The process begins by randomly choosing k objects as the centroids of the k clusters. The objects are assigned to one of the k clusters based on the minimum value of the distance between the object. vector applying Rough Set Concepts to Clustering C= ($C_1, \ldots, C_j, \ldots, C_m$) and the cluster vector X= ($x_1, \ldots, x_j, \ldots, x_n$,). The distance d(C, X)is the distance between the centroid and object. There are many methods for calculation of centroids. The modified and simple centroid calculations for rough sets are then given by:

(i) if $\underline{R}$ (X) $= \varphi$ and $\overline{R}(X) - \underline{R}(X) \neq \varphi$ then $C_j = \frac{\sum_{x \in (\overline{R}(X) - \underline{R}(X))} x_j}{|\overline{R}(X) - \underline{R}(X)|}$

(ii)if $\underline{R}(X) \neq \varphi$then $C_j = \frac{\sum_{x \in \overline{R}(X)} x_j}{|\overline{R}(X)|}$

From the table 1, the performance is Good, $X_{\text{GOOD}} = \{x_1,x_2,x_3,x_4\}$then

$\underline{R}X_{\text{Good}} = \{ x \in U /[x]_R \subseteq X_{\text{GOOD}}\} = \{x_2,x_3,x_4\}$

$\overline{R}X_{\text{Good}} = \{ x \in U /[x]_R \cap X_{\text{GOOD}} \neq \Phi \} = \{x_1,x_2,x_3,x_4x_9\}$

the performance is fair , $X_{\text{Fair}} = \{x_5,x_6,x_7\}$then

$\underline{R}X_{\text{Fair}} = \{ x \in U/[x]_R \subseteq X_{\text{Fair}}\} = \{x_5, x_6, x_7\}$

$\overline{R}X_{\text{Fair}} = \{ x \in U/[x]_R \cap X_{\text{Fair}} \neq \Phi \} = \{x_5, x_6, x_7\}$

the performance is Bad, $X_{\text{Bad}} = \{x_8, x_9, x_{10}\}$then

$\underline{R}\, X_{\text{Bad}} = \{ x \in U/[x]_R \subseteq X_{\text{Bad}}\} = \{x_8, x_{10},\}$

$\overline{R}X_{\text{Bad}} = \{ x \in U/[x]_R \cap X_{\text{Bad}} \neq \Phi \} = \{x_1, x_8, x_9, x_{10}\}$

Let C_1, C_2, C_3 be the centroid of the decision attribute characters Good, Fair, bad respectively. for example if we choose two attributes Age and Eagerness to learn

Table.2. Values for table 1

U	Age(a_1)	Eagerness to Learn (a_2)
x_1	3	1
x_2	1	1
x_3	2	1
x_4	1	2
x_5	2	3
x_6	3	2
x_7	1	3
x_8	3	3
x_9	3	1
x_{10}	3	3

Table.3. Eagemess to learn

C	Age (a_1)	Eagerness to Learn (a_2)
C_1	2	1.2
C_2	2	2.6
C_3	3	2

For age, we get the centroids

$C_1 = (x_1 + x_2 + x_3 + x_4 + x_9)/5 = 2$

$C_2 = (x_5 + x_6 + x_7)/3 = 2$

$C_3 = (x_1 + x_8 + x_9 + x_{10})/4. = 3$.etc.

Calculate d(C,X)

The distance between centroid and each object is calculated as follows. From table 2 and 3

$C_1 = (2, 1.2)$, $x_1 = (3,1)$,

$$d(C_1, x_1) = \sqrt{1^2 + (0.2)^2} = 1.02$$

we get, $d(C_1, x_2) = 1.02, d(C_1, x_3) = 0.2,$

$d(C_1, x_4) = 1.28, d(C_1, x_5) = 1.8, d(C_1, x_6) = 2.05,$

$d(C_1, x_7) = 2.05, d(C_1, x_8) = 1.8, d(C_1, x_9) = 1.02, d(C_1, x_{10}) = 1.8$

$d(C_2, x_1) = 1.8, d(C_2, x_2) = 1.8, d(C_2, x_3) = 1.6, d(C_2, x_4) = 1.16,$

$d(C_2, x_5) = 0.4, d(C_2, x_6) = 1.16, d(C_2, x_7) = 1.07,$

$d(C_2, x_8) = 0.4, d(C_2, x_9) = 1.8, d(C_2, x_{10}) = 0.4,$

$d(C_3, x_1) = 1, d(C_3, x_2) = 2.23, d(C_3, x_3) = 1.4, d(C_3, x_4) = 2,$

$d(C_3, x_5) = 1.41, d(C_3, x_6) = 1, d(C_3, x_7) = 2.23,$

$d(C_3, x_8) = 1.41, d(C_3, x_9) = 1, d(C_3, x_{10}) = 1.41$

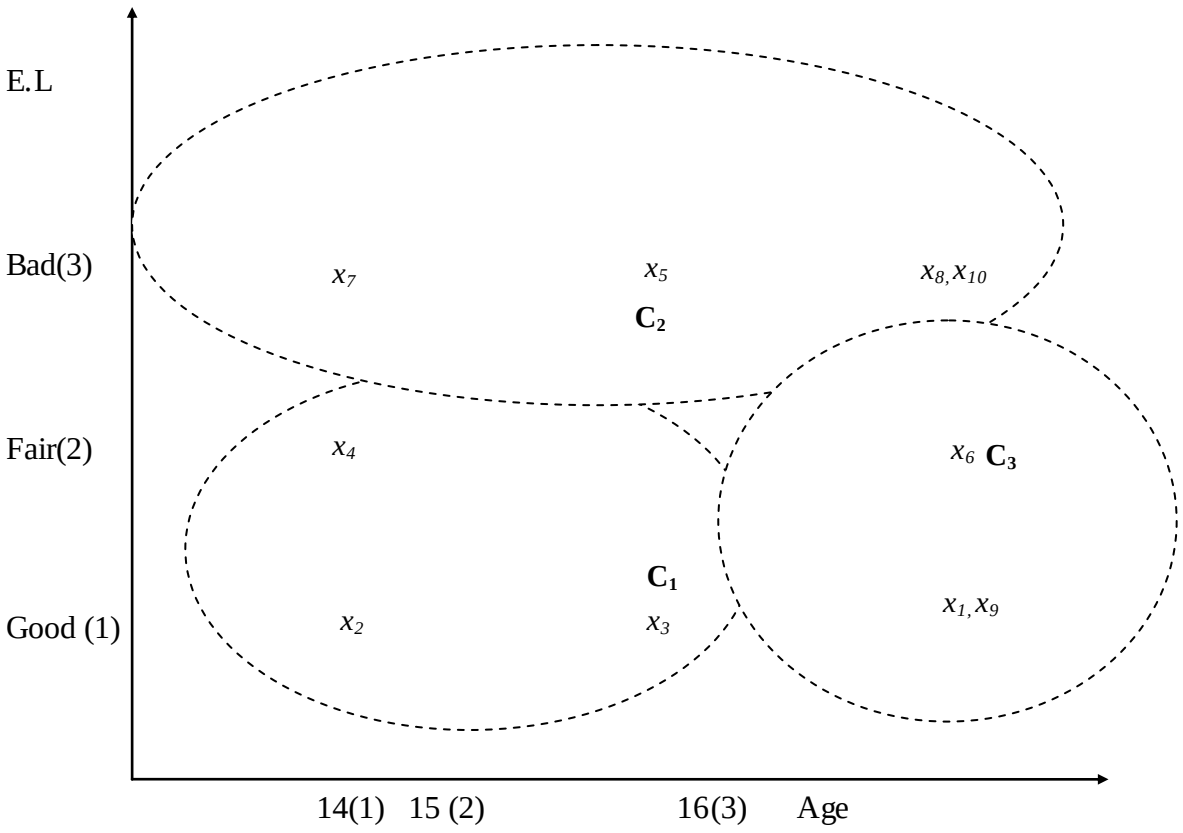

Figure.2. Rough clustering results

Based on the minimum distance between the object and the centroids we have to cluster the objects. See in Figure 2.

5. REFERENCES

[1]. Herstein, I.N., Topics in Algebra, 2nd edition, Wiley, New York, 1975.

[2]. Pawlak, Z. Rough set theory and its applications, Journal of tele communications and information technology, 3/2002.

[3]. Pawlak, Z. Rough Sets, Institute of Theoretical and Applied Informatics, Polish Academy of Sciences, ul. Bałtycka 5, 44 100 Gliwice, Poland University of Information Technology and Management ul. Newelska 6, 01-447 Warsaw, Poland.

[4]. Pawlak, Z. Rough Sets and Data Mining ,Institute of Theoretical and Applied Informatics, Polish Academy of Sciences, ul. Baltycka 5, 44 100 Gliwice, Poland.

[5]. Pawan Lingras and Georg Peter, Applying Rough Set Concepts to Clustering G. Peters et al. (eds.), Rough Sets: Selected Methods and Applications in Management and Engineering, Advanced Information and Knowledge Processing, DOI 10.1007/978-1- 4471-2760-4_2, © Springer-Verlag London Limited 2012