Volume 4, Issue 12, December 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Covering Based Rough Set using K-means

Manisha Modak, Niharika Singhal, Venkatesan M.
School of Computer Science and Engineering VIT University
Vellore, Tamil Nadu, India

Abstract— In information systems, uncertainty and incompleteness of knowledge are general phenomena. To deal with such uncertainty problems, researchers have proposed several methods such as fuzzy theory and rough set theory. Rough set theory is a powerful tool for dealing with vagueness, and incompleteness of knowledge in information systems. But the properties of rough sets as defined by Pawlak does not verify all the properties of rough set where the core idea to separate the distinguishable items from the indistinguishable ones. Also another soft computing k-means algorithm based on rough set recently introduced by Lingras describes the core concept of rough sets, the lower and upper approximation of a set that can be used in clustering. Due to problems existing in the algorithm, a refined rough k-means was proposed. This paper presents generalizations of covering based rough sets to incomplete information systems, involving imprecise observations of attributes. Here we analyse their algorithm with respect to its numerical stability and the stability of the clusters required for the covering. Taking these properties under considerations, the second type of covering based lower and upper approximation operations is used.

Keywords— clustering, rough set, covering, approximations, k-means

I. INTRODUCTION

Over a wide mixture of fields, information are constantly gathered and collected at a sensational pace, particularly at the time of web. But the information found on the surrounding world is often imprecise, incomplete or uncertain. There is quite need of managing such fragmented data in characterization, idea detailing and information examination to draw some conclusion. Tools, which ended up being sufficiently adequate for the investigation of different types of data, particularly, at the point when managing inaccurate, indeterminate or inexact knowledge, are the fuzzy set, granular computing and the rough set theories. Rough sets and fuzzy sets encapsulates two unique feature of defects in knowledge: indiscernibility and unclearness

Zdzislaw Pawlak [3] developed Rough set theory in the early 1980's. It is a current mathematical approach to data mining and data analysis. A rough congregate is a group, whose components can be a piece of one or more groups. The lower and upper estimates of the rough set are focused on identicalness relations on a set. The lower estimate of a rough cluster contains elements that just belongs to that cluster. The Upper estimate of a rough cluster holds elements of the cluster which are also the part of other clusters. The paramount part of rough set theory is partition or equivalence relation as presented by Pawlak [3]. It serves as the mathematical ground for the rough set theory. The category of equivalence classes is called as a partition of the universe [7].

The benefit of the rough set method is that it does not require any additional information regarding the data, like Probability in statistics or membership present in fuzzy set theory. Because of the advantage that involves rid of demand of excess information from data and its power to draw out rules straight from the data itself has increased its usage in many domains The rough set methodology is focused around the preface that bringing down the level of accuracy in the information makes the data pattern more visible which are exhibited in the form of assortment or decision rules developed from a set of instance. Different real life-purposes of rough set theory have demonstrated its utility in numerous areas. Very likely new domain of application of the rough set concept seems to rise within a brief period of time. They incorporate rough control, rough data bases, rough information retrieval, rough neural network and others.

Also the basic objective of clustering is the procedure of classifying the data into classes in which the object that belong to one group have high resemblance in equivalence to others. Clustering is also an unsupervised learning. The employment of clustering in mining is ability to handle the noisy data, different types of attributes, order of input records, scalability, and discovery of clusters with arbitrary shape, high dimensionality and constraint based clustering. Clustering can be classified as partitioning based, hierarchical based, density based, grid based and model based methods. Partitioning based clustering methods divides the n objects in k partitions such that each object essentially holds at any rate one object and all objects must belong to precisely one group. Hierarchical based clustering methods creates a graded decomposition of the data objects. Density based clustering methods classify the object based on the distance metrics. Grid based clustering method divides object into finite number of clusters such that the processing time decreases. Model based methods classify the data objects based on the density function which represents the dispersion of objects into the space. All the clustering algorithms are suitable for their respective situation. The most frequently used clustering is k means algorithm in which partition object according to the predefined clusters .The cluster are formed to get an optimal value for the objective partitioning criterion which is also known as similarity function in which similar objects are put in one cluster and away from the dissimilar objects.

© 2014, IJARCSSE All Rights Reserved

However, rough set is confined to numerous real-world applications. To get over this restriction, there are two primary techniques to extrapolate the standard rough sets. One systematic way is to widen the concept of equivalence relation to other binary relations. The other significant method is to substitute a partition of the universe with a covering. Thus covering based rough set is an elongation to the rough set and gives more precise answer when handling with incertitude data and vagueness. There exists three subdivision of covering based rough set. The ideas used in calculating the lower estimate are alike but the ideas of creating upper approximations are unlike in nature. Thus, the idea of second type of covering based rough set is favourable as it has similar properties as that of rough set.

This paper presents the classical k-means clustering algorithm along with the introduction of Lingras rough k-means in which the covering based rough sets are generalized. Taking into consideration the objective function, numerical stability and the stability of the clusters, the second type of covering based lower and upper approximation operations are used.

II. BACKGROUND

A. Rough Set Theory

Pawlak's proposed a theory called the basic rough set theory. This theory takes some prerequisite presumption because of which it has been increased in all dimensions. Rough set are often induced by the description of the objects based on the precise observation of an insufficient number of the attributes. Some additions are made to the rough set theory and a theory named Covering based rough set is developed. Rough set provide the mathematical tool to find the shrouded examples in data systems. It is additionally used to distinguish the incomplete or complete dependencies in information and furthermore take out redundancies. Indiscernible Relation or Equivalence relation and Approximation regions are the two concepts that can be used. Indiscernible relation is used to fragment the universe into equivalence classes. . The inability to compare two or more values due to the lack of exactness in measurement is referred as indiscernibility.

B. Lower and Upper Approximation

Calculation of the uncertainty can be done with the help of the mathematical tool called Rough set. Two estimates namely upper and the lower are used to define a rough cluster. Upper estimate is the superset of the lower estimate. The object which belong to the lower estimate definitely lie in that particular cluster. These objects do not lie in the any other cluster. Whereas the object which belong to the upper estimate do not share such surety. These object must belong to the at least one more upper estimation of any other cluster. Roughness can be measured as the intersection of the upper estimate and the lower estimate.

Let U be the non-empty collection of objects known as the universe of discourse and equivalence relation over U is named as R. For a set which is the subset of U and the approximation space which is defined as A= (U, R) the lower estimation of Y under the relation R and in the approximate space A is formulated as

$$\underline{R}Y = \{y \in U \mid [Y]_R \subseteq Y\}$$

And an Upper estimation of Y in A under R is defined as

$$\overline{R}Y = \{\, y \in U \mid [Y]_R \cap Y \neq \Phi\}$$

Consider the following information system

Table 1: Information System Table

objects	S1	S2	S3	S4	S5
A1	**1**	**2**	**1**	**2**	**2**
A2	**1**	**1**	**0**	**1**	**0**
A3	**2**	**2**	**1**	**1**	**0**
A4	**2**	**1**	**1**	**1**	**0**
A5	**1**	**1**	**0**	**1**	**0**
A6	**2**	**2**	**1**	**0**	**1**
A7	**2**	**2**	**2**	**2**	**2**
A8	**2**	**2**	**1**	**0**	**1**
A9	**2**	**1**	**1**	**1**	**0**

From the above table
Equivalent classes

{{A1}, {A5, A2}, {A3}, {A4, A9}, {A6, A8}, {A7}}

© 2014, IJARCSSE All Rights Reserved

The Target Class

$$\{A1, A2, A4, A5, A9, A10\}$$

Lower approximation Region for the above equivalence classes

$$\underline{RY} = \{\{A1\}, \{A5, A2\}, \{A4, A9\}\}$$

Upper approximation region for the above equivalence classes

$$\overline{RY} = \{\{A1\}, \{A5, A2\}, \{A4, A7, A9\}\}$$

Roughness according to rough sets

$$\overline{RY}\text{-}\underline{RY} = \{A7\}$$

C. Covering Based rough Set

Covering Based rough set concept requires some basic construct such as reducible element and neighborhood function. After Zakowski broadened the concept of the rough set defined by Pawlak [11, 13] to covering-based rough set, several other type of covering based rough set are formulated. Covering based uses covering instead of partitions. Covering approximation space is formulated as (U, C) .

Considering the example given above in Fig 2.1, let us assume a covering set

$$C = \begin{Bmatrix} \{A1, A5, A3\}, \{A2\}, \{A2, A5, A3\}, \\ \{A4, A7, A8\}, \{A7\}, \{A6\} \end{Bmatrix}$$

From the above covering C

$$k1 = \{A1, A5, A3\},\ k2 = \{A2\}, k3 = \{A2, A5, A3\},\ k4 = \{A4, A7, A8\},\ k5 = \{A7\},\ k6 = \{A6\}$$

The Target Class = $\{A1, A2, A4, A5, A9, A10\}$

Lower approximation Region for the above equivalence classes

$$\{\{A2\}\} = \{A2\}$$

Upper approximation Region for the above equivalence classes

$$\{\{A1, A3, A5\} \cup \{A2\} \cup \{A2, A5, A3\}\} = \{A1, A2, A3, A5\}$$

Roughness according to covering based rough sets

$$\overline{RX}\text{-}\underline{RX} = \{A1, A3, A5\}$$

Definition 2.1

Let U be space to be taken into consideration called the universe of discourse and C be the collection of subsets of U. C is the cover of U if subset of C is non-empty and union of covering set in the covering C form the universe of discourse denoted as $\cup\, C = U$. If C is cover of U then only the covering approximation space can be defined as an ordered pair of (U, C). As every partition of U is also the covering of U, it can be concluded that covering is the broadened concept of a partition.

2.1 Second Type of Covering

Zhu delineated the upper estimation operator for covering based rough set. The parameter in which the second type of covering based rough set differs from the first type in the upper estimation whereas the lower estimation parameter is same for both the former and the latter. The second type of upper estimation is defined as

$$\overline{Y} = \cup \{K | K \in C, K \cap Y \neq \phi\}$$

The upper and lower estimation are the concepts which are utilized both in rough set and covering based rough set. The second type of covering based rough set uses the concept of covering to calculate the lower and the upper estimation for the rough set. Even the value of the parameter roughness which is determined as the intersection of the lower and the upper estimation correspond to more elements in covering based rough set than the normal rough sets. So covering based rough set is more preferable than normal rough sets.

III. LITERATURE SURVEY

A central premise of the conjecture of rough sets is that objects is characterized by an equivalence relation, or identically a partition of the universe [1, 9]. Starting from the time that the presentation of rough sets in the beginning 1980s, numerous endeavours was made to evacuate this supposition. By the way that a covering is an abstraction of a partition, it is veritable to try to understand covering based induction of rough sets. Rough set theory has been proposed by Pawlak [1] as a means for conducting relation with the vagueness. In the Pawlak rough set model, an equality connection offers two extra proportionate structures. The rough set concept converges to some degree with numerous other mathematical tools formulated to address with unclearness and dubiety, in specific with the Dempster-Shafer hypothesis of evidence

© 2014, IJARCSSE All Rights Reserved

[1]. The primary deviation is that the Dempster-Shafer theory employs belief function as the chief tool, while rough set theory makes utilization of Upper and Lower estimates. Some other relationship exists amongst fuzzy set and rough set theory

[1, 2]. One is the partition actuated by the equivalence relation, while the other is the atomic Boolean algebra with the partitions as its set of items.

In 1983, Zakowski [13] first suggested the concept of covering based rough set estimates. However, dissimilar to the Pawlak estimate operators, the generalized approximation operators are no longer dual to each other with regard to determine complement [6]. William Zhu- delineated the holdings of rough sets to bump out the factors beneath which the second type covering based rough sets will fulfill these holdings. Jianguo Tang, She Kun, Zhu William - advised that in the first and third type covering based rough sets, the lower estimate area gives greater degree of data than primary lower estimate regions and upper estimate regions gives less amount of data considering the original upper estimate in some particular situations. ChengyiYu, Fan Min, Zhu William [5, 11] delineated the unrestrained matroidal structure of covering generalized rough sets and the relationships amongst the reducible basics of covering based rough sets to the reducible matroid theory. Jianguo Tang, Kun She, William Zhu – delineated the elaboration of covering based rough sets by bringing down the size of the covering elements, such that the object appears more accurate [5].

IV. ALGORITHMS USED

4.1 Classical K-means

The classical *k*-means algorithm takes the input parameter, *k*, and partitions a set of *n* objects into *k* clusters so that the resultant similarity within the cluster is high but similarity among the clusters is low.

1. The algorithm randomly chooses k points as the initial cluster centres.
2. Each point in the dataset is allotted to the nearest cluster, depending upon the Euclidean distance between each point and each cluster center.
3. Each cluster centre is recalculated as the mean of the points in that cluster.
4. Repeat Steps 2 and 3 until the clusters come together or becomes adjacent

4.2 Rough K-means

Sometimes there exists inadequate knowledge to exactly delineate the clusters as sets. Then comes the usage of rough sets. Few of the standard properties of the rough set are:

i. A data object can be a portion of at most one lower approximation.
ii. A data object which belongs to lower approximation of a set is also a part of its corresponding upper approximation.
iii. If a data object does not belong to any lower approximation, it implies that it belongs to two or more upper approximations.

The Lingras rough K-means algorithm can be put forward as follows:

1. Choose an initial group of n objects into k clusters.
2. Allocate each object to the Lower bound (L(c)) or upper bound (U(c)) of cluster/ clusters respectively as:

For each object a, let $d(a, ci)$ be the distance within itself and the centroid of cluster ci. The difference between $d(a, ci) - d(a, cj)$, $1 \leq i, j \leq k$ is utilized to regulate the membership of a as follows:

- If $d(a\ ci) - d(a, cj) \leq$ threshold, then $a \in U(C_i)$ & $a \in U(C_j)$. Furthermore a will not be a part of any lower bound
- Otherwise, $v \in L(C_i)$,such that $d(a, xi)$ is the minimum for $1 \leq i \leq k$. In addition, $a \in U(C_i)$.

3. For each cluster ci re-compute center in accordance to the following equations the weighted sequence of the data points in its lower bound and upper bound.

$$C_{i=} \begin{cases} W_{lower} \times \frac{\sum_{a \in L(c)} A_j}{|L(c)|} + W_{upper} \times \frac{\sum_{a \in U(c)-L(c)} A_j}{|U(c) - L(c)|} \\ \qquad if\ |U(c) - L(c) \neq \phi| \\ W_{lower} \times \frac{\sum_{a \in L(x)} A_j}{|L(C)|} \quad otherwise \end{cases}$$

Where $1 \leq j \leq k$. The parameters W_{lower} and W_{upper} agrees to the relative importance of lower and upper bounds.

If the overlap measure is satisfied, i.e. cluster centers are similar to those in former iteration, then stop; else go to step2

The concept of lower and upper approximations in Rough k means algorithm deals with two weight argument *i.e.* W_{lower} and W_{upper}.These arguments delineate the upper and the lower bound of the rough set. As the measure of W_{upper} gains, the root mean square error also grows. The increment in the error can be easily ended as the elements in the lower bound

© 2014, IJARCSSE All Rights Reserved

of the cluster exactly belongs to the cluster which on the contrary is inevitable for the elements of the upper bound. These element may or may not belong to the cluster.

4.3 Limitation of Rough K-means

Rough K-means clustering algorithm has some limitations that needs to be modified. The following limitations are:

1. Number of clusters: For clustering one has to determine the number of cluster prerequisitely. This task is the challenging task in the algorithm as the number of cluster depends upon elements in the dataset.
2. Outliers: Outliers are the object which do not belong to any cluster in the system. When the outlier is present in the dataset then the iterations of the algorithm do not produce a stable result. Moreover, the sum of squared errors also increases.
3. Numerical stability: Algorithm is not numerically stable as it can sometimes result in empty clusters. Empty cluster occurs when no points in the dataset belong to a particular cluster.
4. No globular shapes and sizes: When the clusters are of different density and non-globular shapes, then the result which is produced by the algorithm is not optimal. Rough k means always form the convex shape clusters.

V. PROPOSED COVERING BASED ROUGH K-MEANS ALGORITHM

This algorithm works for covering based rough set data with the application of Lingra's algorithm with some modifications. The algorithm is as follows:

1. Choose an initial group of n objects into k clusters and divide them into definite number of covers, c_i such that $c_1 \cup c_2 \cup \ldots \cup c_n = C$
2. Calculate their lower and upper approximations using the formula for Second type of Covering.
3. For each cluster ci re-compute center in accordance to the following equations the weighted sequence of the data points in its lower bound and upper bound.

$$C_{i=} \begin{cases} W_{lower} \times \dfrac{\sum_{a \in \underline{R}(C)} A_j}{\left|\underline{R}(C)\right|} + W_{upper} \times \dfrac{\sum_{a \in \overline{R}(C) - \underline{R}(C)} A_j}{\left|\overline{R}(C) - \underline{R}(C)\right|} \\ \qquad if \left|\overline{R}(C) - \underline{R}(C) \neq \phi\right| \\ W_{lower} \times \dfrac{\sum_{a \in L(x)} A_j}{\left|\underline{R}(C)\right|} \quad otherwise \end{cases}$$

4. Allocate each object to the Lower bound $\underline{R}(C)$ or upper bound $\overline{R}(C)$ of cluster/ clusters respectively as:

For each object a, let d(a, ci) be the distance within itself and the centroid of cluster ci. The difference between d(a, ci)− d(a, cj), 1≤ i, j ≤ k is utilized to regulate the membership of a as follows:

- If d(a ci)− d(a, cj) ≤ threshold, then $a \in U(C_i)$ & $a \in U(C_j)$. Furthermore a will not be a part of any lower bound
- Otherwise, v∈ L(C_i),such that d(a, xi) is the minimum for 1≤ i ≤ k. In addition, a∈ U (C_i).

5. Again compute the centroid using the above formula and repeat the steps 3 and 4 until the algorithm converges

VI. IMPLEMENTATION DETAILS

A dataset containing 4 objects is considered as training data points object each having two attributes namely weight index and pH. Each attribute is denoted as the coordinate of each object. Presuming as the number of clusters to be two (k=2) it is being determined which object belongs to either of the clusters.

Table. 2. Object Data Set

Object	Attribute1 (x): weight	Attribute2 (y) : pH value
A	**1**	**1**
B	**2**	**1**
C	**4**	**3**
D	**5**	**4**

© 2014, IJARCSSE All Rights Reserved

6.1 K-means Algorithm

Iteration 1:

- Suppose we use object A and B as the first centroids.
- Let $C_1$and C_2 refers to the coordinate of the centroids, then C_1=(1,1) and C_2=(2,1)

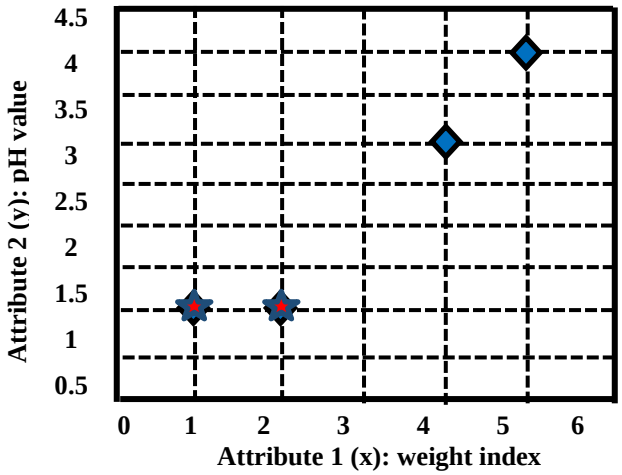

Fig. 1. K-means iteration 0

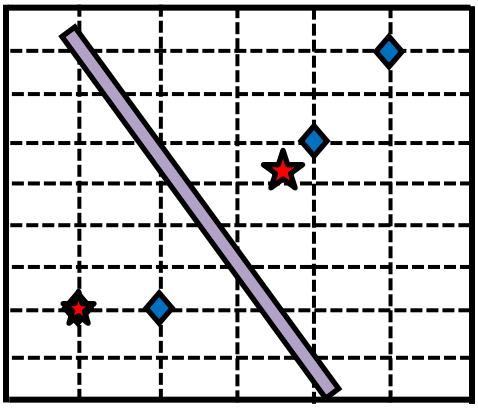

Fig. 2. K-means iteration 1

- **Distance between Objects and Centroids:** the distance between cluster centroid to each object is calculated using Euclidean Distance. The distance matrix at iteration 0 is

$$D^0 = \begin{bmatrix} 0 & 1 & 3.61 & 5 \\ 1 & 0 & 2.83 & 4.24 \end{bmatrix} \quad c_1 = (1,1) \text{ Cluster-1}$$

$$c_2 = (2,1) \text{ Cluster-2}$$

- Each column in the distance matrix represents an object.
- The first row of the distance matrix refers to the distance between each object to the centroid of first cluster and the second row shows the distance of each object to the centroid of second centroid.
- As for example, distance from object C = (4, 3) to the first centroid $c_1 = (1,1)$ is $\sqrt{(4-1)^2 + (3-1)^2} = 3.61$ and its distance from the second centroid $c_2 = (2,1)$ is $\sqrt{(4-2)^2 + (3-1)^2} = 2.83$

Iteration 2:

- **Clustering of Objects***:* Each object is assigned based on the minimum distance.
- Object A is placed in cluster 1, object B in cluster 2, object C in cluster 2 and object D in cluster 2.
- The elements of Group matrix below is 1 if only when the object is assigned to that specific cluster.

$$G^0 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \end{bmatrix}$$

- Now the distance of all objects from the new centroids is calculated.
- In the similar manner distance matrix calculated is

$$D^1 = \begin{bmatrix} 0 & 1 & 3.61 & 5 \\ 3.14 & 2.36 & 0.47 & 1.89 \end{bmatrix}$$

$c_1 = (1,1)$ Cluster-1, $c_2 = \left(\frac{11}{3}, \frac{8}{3}\right)$ Cluster-2

- On the basis new distance matrix, we move the object B to Cluster 1 and all other remaining objects stays the same. The Group matrix is shown below

$$G^1 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix}$$

- Same calculation is done to calculate the new centroids coordinate depending on the clustering of previous iteration.

- Cluster 1 and Cluster 2 both has two objects each, thus the new centroids are calculated as

$$c_1 = \left(\frac{1+2}{2}, \frac{1+1}{2}\right) = \left(\frac{3}{2}, 1\right) \quad c_2 = \left(\frac{4+5}{2}, \frac{3+4}{2}\right) = \left(\frac{9}{2}, \frac{7}{2}\right)$$

© 2014, IJARCSSE All Rights Reserved

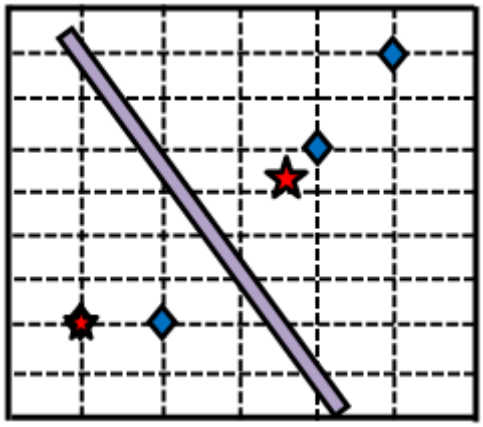

Fig. 3. K-means iteration 2

- The new distance matrix calculated is:

$D^2 = \begin{bmatrix} 0.5 & 0.5 & 3.20 & 4.61 \\ 4.30 & 3.54 & 0.71 & 0.71 \end{bmatrix}$ $c_1 = \left(\frac{3}{2}, 1\right)$ Cluster-1, $c_2 = \left(\frac{9}{2}, \frac{7}{2}\right)$ Cluster-2

- Re-assigning each object based on the minimum distance the group matrix turns out to be

$$G^2 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix}$$

- The results obtained from the above iterations shows that G^2=G^1 which reveals that the objects does make any further movements anywhere in the group.

Final clustering results is as follows**:**

Table. 3. Grouping of objects into fixed clusters

Object	Feature1(x): weight index	Feature2 (y): pH value	Cluster Results
A	1	1	1
B	2	1	1
C	4	3	2
D	5	4	2

6.2 Rough K-means

Considering the same dataset for rough k-means algorithm with the value of k=2 the following is obtained.

Step 1: Chose the initial cluster of n objects into k objects where n=4 and k=2 according to the given dataset.

Step 2: Initially consider the cluster centers as $C_1(1,1)$and $C_2(3,4)$ and a threshold value of 0.9. The value of $\boldsymbol{W_{lower}}$ and $\boldsymbol{W_{upper}}$ is taken as 0.3 and 0.7 respectively.

Step 3: Using the Euclidean distance as that of classical k-means, the distance of the data point from the given cluster centers is calculated.

Table. 4. Distance of each objects from the centers of both the clusters

Objects	Centroid C_1	Centroid C_2
A	0	1
B	1	0
C	3.61	2.83
D	5	4.24

Step 4: Calculations of the upper and the lower approximations of the rough set for the prescribed data value along with the centroids for each iteration until the algorithm converges.

© 2014, IJARCSSE All Rights Reserved

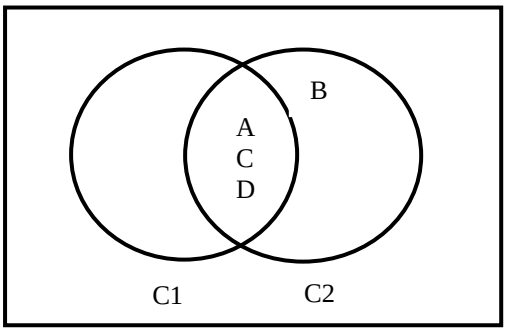

Fig. 4. Position of objects after iteration 0 w.r.t centroids $C_1(1,1)$ and $C_2(2,1)$

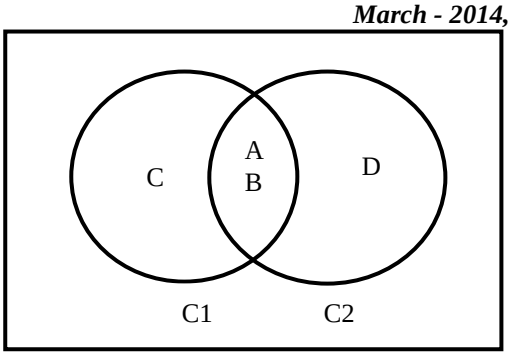

Fig. 5. Position of objects after iteration 1 w.r.t centroids $C_1(2.3,1.8)$ and $C_2(2.9,2.1)$

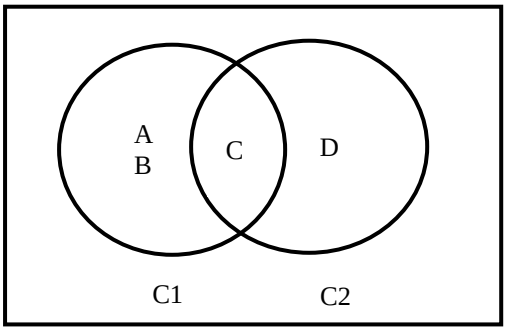

Fig. 6. Position of objects after iteration 2 w.r.t centroids $C_1(3.3, 2.3)$ and $C_2(3.6,3.3)$

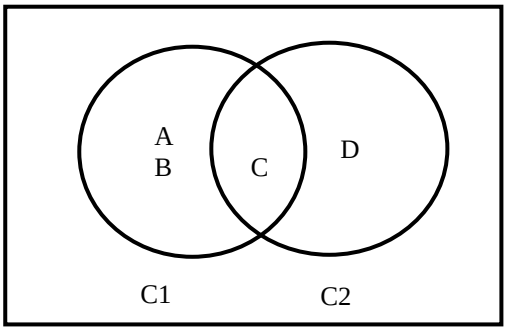

Fig. 7. Position of objects after iteration 3 w.r.t centroids $C_1(3.25,2.55)$ and $C_2(4.3,3.3)$

After the completion of third iteration, the clusters concludes the same approximation results and hence the algorithm converges.

The Rough K-means is then tested for various threshold values of 0.1, 0.5, 0.8 and 0.9. It has been found that the number of iterations in which the algorithm converges is shown in the graph below.

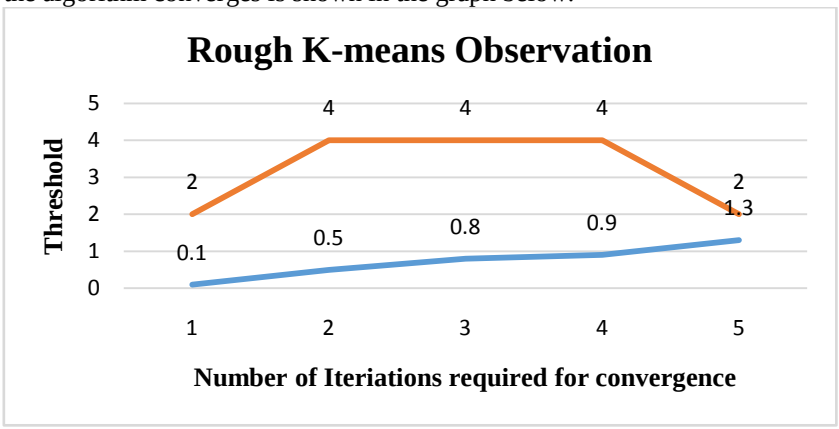

Fig. 8. Analysis of Rough K-means with various values of threshold

6.3 Rough K-means using Covering based data

Consider the data set

Table 5. Dataset for covering based data

Object	A	B
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

© 2014, IJARCSSE All Rights Reserved

Let U be the universe containing the following set of objects {1,2,3,4,5,6,7} and C be the covering of the above data set (*i.e.* universe) where

$$C = \{\{1,3,5\},\{1,2\},\{5,6,7\},\{4,6\},\{2\}\}$$

Here $c1 = \{3,5,7\}, c2 = \{1,2\}, c3 = \{5,6,7\}, c4 = \{4,6\}, c5 = \{2\}$

Such that $c1 \cup c2 \cup c3 \cup c4 \cup c5 = C$

Lower approximation is $\underline{R}(C) = \{4,6\}$

Upper approximation is

$$\overline{R}(C) = \{\{3,5,7\},\{1,2\},\{5,6,7\},\{4,6\}\} = \{1,2,3,4,5,6,7\}$$

Boundary region approximation

$$B(C) = \overline{R}(C) - \underline{R}(C) = \{1,2,3,5,7\}$$

The above results gives the position of the objects after iteration 0

Consider the values of W_{upper} , W_{lower} and Threshold to be 0.7, 0.3 and 0.9 respectively. Calculating the centroids of the two clusters using the formula of step 3 of Rough K-means algorithm the values obtained are

$$C_1 = (3.105, 4.11) \text{ and } C_2 = (3.105, 4.11)$$

Now applying rest of the Rough k-means algorithm steps explained above in the following data sets gives the following results shown below.

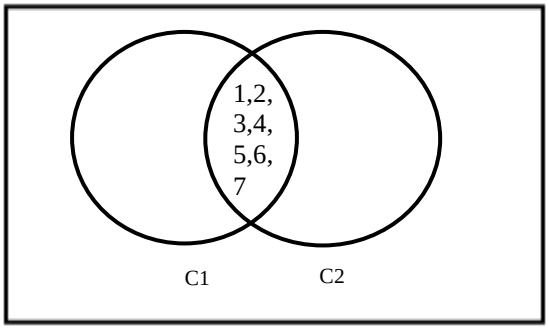

Fig. 9. Position of objects after iteration 2 w.r.t centroids $C_1(3.105, 4.11)$ and $C_2(3.105, 4.11)$

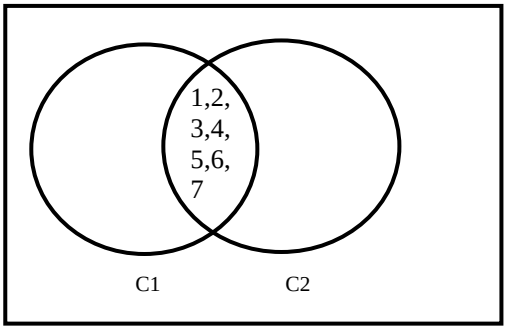

Fig. 10. Position of objects after iteration 2 w.r.t centroids $C_1(3.105, 4.11)$ and

The above figures show that the all the coverings belongs to upper approximation of both cluster 1 and 2 respectively. Thus by using covering, the algorithm converges faster than the Rough K-means

VII. CONCLUSION

This paper analyzed the dataset with respect to conventional k means developed by Pawlak and rough k means algorithm proposed by Lingras. Through the observations it can be concluded if the datasets deal with the precise data that the k means algorithm works fine. But information in the system is generally incomplete. To handle such situations one requires rough set theory. This leads to the development of rough k-means algorithm which takes into consideration the vague data. Moreover, a challenge still exists in the Rough k means algorithm which is not solved yet that is with the selection of the initial parameters namely the weight parameters and the threshold value.

Despite the limitations, the observations in the implementation concludes that the threshold value of 0.8 leads to the faster convergence of the algorithm. Thus, to overcome these limitations, the concept of second type of covering based rough set was introduced. The magnitude of roughness is greater in covering based rough is greater than the normal rough set. Thus the overlapping in case of covering based rough set is more than the normal rough set, the former is more applicable for the real life situations.

REFERENCES

[1] Pawlak, Z., and Skowron, A. Rough membership functions. In R. Yaeger, M. Fedrizzi, and J. Kacprzyk, Eds. Advances in the Dempster-Shafer Theory of Evidence. Wiley, New York, 1994, 251—271

[2] Skowron, A., and Grzymala-Busse, J. W. From rough set theory to evidence theory. In Advances in the Dempster Shafer Theory of Evidence. R. R. Yaeger, M. Fedrizzi, and J. Kacprzyk, Eds. Wiley, New York, 1994, 193--236.

[3] Z. Pawlak, Rough Sets: Theoretical Aspects of Reasoning about Data, Kluwer Academic Publishers, Boston, 1992

[4] P. Samanta, M.K. Chakraborty, Covering based approaches to rough sets and implication lattices,in: Proceedings of the Twelfth International Conference on Rough Sets, Fuzzy Sets, Data Mining and Granular Computing, LNCS(LNAI) 5908, 2009, pp. 127–134.

© 2014, IJARCSSE All Rights Reserved

[5] W. Zhu, "Properties of the Second Type of Covering-Based Rough Sets," Proc. First Int'l Workshop Granular Computing and Brain Informatics (GrC & BI '06), IEEE Int'l Conf. Web Intelligence (WI '06), pp. 494-497, Dec. 2006.

[6] Pawlak, Z, 1991, „Rough sets - Theoretical aspects of reasoning about data", Dordrecht:Kluwer Academic Publishers, pp. 68-162.

[7] P. Lingras, C. West, Interval set clustering of web users with rough *k*-means, J. Intell. Inform. Syst. 23 (2004) 5–16.

[8] R. Jensen, Q. Shen, Semantics-preserving dimensionality reduction: rough and fuzzy-rough-based approaches, IEEE Transactions on Knowledge and Data Engineering 16 (12) (2004)

[9] William Zhu, Member, IEEE, and Fei-Yue Wang, Fellow, IEEE, "On Three Types of Covering-Based Rough Sets", IEEE transactions on knowledge and data engineering, VOL. 19, NO. 8, AUGUST 2007

[10] Z. Pawlak, Rough Sets, International Journal of Computer and Information Sciences 11 (5) (1982) 341–356.

[11] W. Zhu and F. Y. Wang, "Properties of the first of covering based rough sets", sixth IEEE International Conference on Data-Mining Workshop (ICDMW'06), March – 2006.

[12] E. Tsang, D. Cheng, J. Lee and D. Yeung, "On the upper approximations of covering generalized rough sets," in Proc. 3 rd International Conf. Machine Learning and Cybermetics, 2004, pp. 4200-4230.

[13] B. K. Tripathy and G. K. Panda, "On covering based approximations of classification of sets", B.C. Chien et al (Eds) IEA/AIE 2009, springer-Verlag, Berlin, Heidel berg, pp. 777-780, 2009.

[14] A. An, Y. Huang, X. Huang, N. Cercone, Feature Selection with Rough Sets for Web Page Classification Transactions on Rough Sets, SpringerLink Publishers, 2004, pp. 1–13

[15] N. Zhong, J.Z. Dong, and S. Ohsuga, "Using Rough Sets with Heuristics to Feature Selection," J. Intelligent Information Systems, vol. 16, no. 3, pp. 199-214, 2001

© 2014, IJARCSSE All Rights Reserved