# RCM OVER CLOUD FOR DATA MINING

**MUKUNDKUMAR JHA, DEEPAK KHUSHWAHA, NAWAZ ASLAM, AMARJEET SINGH YUMNAM**

***Abstract*— Today cloud computing is providing a variety of computing resources from servers and storage to enterprise applications such as security, voice and email all is delivered over the Internet. With this growth of World Wide Web there arises a need for finding and analyzing this huge amount of data. Mining over a huge data cost heavily. There are various algorithms for data mining, which could be used over the cloud. Clustering is one form of unsupervised learning used in data mining. We are adapting Rough C-mean algorithm for clustering the data sets and using it for any kind of predictions and decision-making.**

***Index Terms*—Cloud computing; Clustering algorithm; Mining; Rough C-Mean;**

## I. INTRODUCTION

In this era of Internet, information and data are growing continuously. Behind a variety of Internet services and applications, there are normally enormous amounts of data. The amount of information is increasing at a rapid speed. The number of pages grows exponentially over time. Hundred billions even trillions of web indexes exist.

Cloud computing is based on Internet, where resource's, software and hardware are provided to customer on demand. Amazon, Google, Microsoft etc. are some of the companies that are providing cloud services. The cloud-computing model has brought many benefits such as low-cost, fault-tolerant mechanism, the fast computing speed, more convenient program development and so on.

## II. CLOUD COMPUTING SERVICES

There are three type of cloud computing services:

1. Software as a Services (SaaS)
2. Platform as a Services (PaaS)
3. Infrastructure as a Services (IaaS)

### *A. Software as a Services*

A software package cost very high so in cloud the software is provided as a service . Client would be charged on the

amount he uses. In cloud the client is not responsible for managing or controlling the resources. The cloud controls and manages the resources and provides it when customer demands.

### *B. Platform as a Services*

The cloud provider provides a cloud base environment for a complete life cycle of development .The client doesn't have to purchase underlying hardware, software and other resources.

### *C. Infrastructure as a Service*

Hardware and other resources changes over time which cost very high. Cloud Provides computing resources such as server, data center etc. to the client and the provider maintains the resources for the client

## III. CLOUD DEPLOYMENT MODELS

Cloud has four types o deployment models :

- Private cloud
- Public cloud
- Community cloud
- Hybrid cloud

### *A. Private Cloud*

This kind of cloud is own by organization or a group of people, it is also manages by the organization it self. The cloud might or might not be build by the organization.

### *B. Public cloud*

This kind of cloud is owned by a organization but is available for use of public. Though maintained by the organization, its services is available for use to the public at anytime.

### *C. Community cloud*

This kind of cloud is available for use to a small no of users. In this the small organizations which doesn't have much capital but have similar requirements pull together there resources to build a cloud.

### *D. Hybrid Cloud*

Hybrid cloud is a combination of private, public and Community cloud. In this cloud some services may be a private or public or a community. In this each service is a different type of cloud. A cloud will also get some features of other types of cloud.



ISSN: 2278 – 7798 

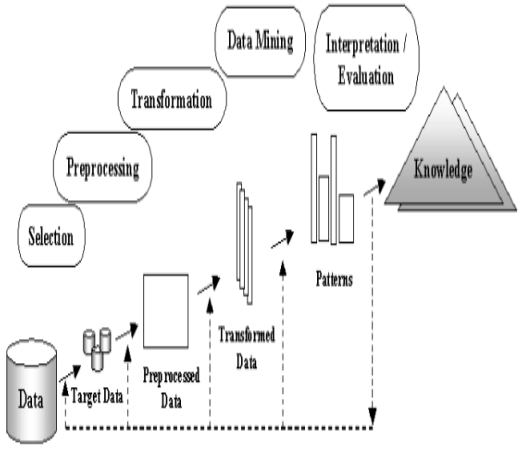
## IV. DATA MINING

Data mining is subfield of computer science. Amount of data available has a vertical growth in this past decade .Analyzing and fining useful information from this huge data has became a must in the fiend of computer science .There has been a lot of research going on in the field of Data mining .Various algorithm has been developed for mining .Data mining can also be extended on web. Three type Web data mining are as follows:

[1] Web structure mining- mining the structure of the web
[2] Web Content mining- For mining the content of the web to gather useful information.
[3] Web log mining- mining the web log to study the usage pattern of web by the users

Mining is done to make prediction and to make decision based on the past .

With this growth of cloud there arises a need for mining the




cloud. Mining consist of various steps as shown[1] in fig. 1.

Fig.1 Data Mining Process

There are various algorithms for data mining, which already exist such as [2]

a) Association Rule Learning Algorithms: Association Rule Learning Algorithms is used to find similar relationship between variable in a large data set.
b) Classification Algorithms: Is a model finding process used for portioning the data into different classes according to some constrains.
c) Clustering Algorithms: Unlike classification, clustering is to enable all records to compose different classes or cluster before we know in advance how many classes the target database has.
d) Stream Data Mining Algorithms: It is the process of extracting knowledge structure from continuous, rapid data records.

In this paper Rough C-Mean algorithm that is also a Clustering Algorithm is being used to demonstrate how mining could be done over the cloud.[4] Rough C-mean is a technique to find the centroid of a cluster from a given data set. The cluster formed using Rough C-mean represents the classifier among the data-set implying that the point or the data-point belongs to which classified cluster. The collection of such information leads to the aggregation of information and forms Knowledge that could be used to predict or make decision.

## V. ROUGH C-MEANS IMPLEMENTATION

Let $\underline{A}(\beta i)$ and $A(\beta i)$ be the lower and upper approximations of cluster $\beta i$, and $B(\beta i) = \{A(\beta i) - \underline{A}(\beta i)\}$ denote the boundary region of cluster $\beta i$. In this algorithm, the concept of c-means algorithm is extended by viewing each cluster $\beta i$as an interval or rough set. Though, it is likely to define a pair of lower and upper bounds $[\underline{A}(\beta i), A(\beta i)]$ or a rough set for every set $\beta i \subseteq U$, U be the set of objects of concern . The family of upper and lower bounds are required to follow some of the basic rough set properties such as:

- An xj can be associated with at most one lower bound;

  $(\beta i) \Rightarrow xj \in A(\beta i)$; and $xj \in A$

- An object xj is not part of any lower bound $\Rightarrow$ xj belongs to two or more upper bounds.

Incorporating rough sets into c-means algorithm, Lingras and West [3] proposed rough c-means algorithm. It add the lower and upper bounds concept into c-means algorithm. It groups the object space into two region - lower approximation and boundary region. The mean (centroid) is calculated based on the weighting average of the lower bound and boundary region. All the objects in lower approximation take the same weight w while all the objects in boundary take another weighting index $\tilde{w}$ $(= 1 - w)$ uniformly. Calculation of the centroid is modified to include the effects of lower as well as upper bounds. The new centroid calculation formula for rough c-means is given by:



ISSN: 2278 – 7798 

$$v_i = \begin{cases} w \times \mathcal{A} + \tilde{w} \times \mathcal{B} & \text{if } \underline{A}(\beta_i) \neq \emptyset, B(\beta_i) \neq \emptyset \\ \mathcal{A} & \text{if } \underline{A}(\beta_i) \neq \emptyset, B(\beta_i) = \emptyset \\ \mathcal{B} & \text{if } \underline{A}(\beta_i) = \emptyset, B(\beta_i) \neq \emptyset \end{cases}$$

$$\mathcal{A} = \frac{1}{|\underline{A}(\beta_i)|} \sum_{x_j \in \underline{A}(\beta_i)} x_j; \text{ and } \mathcal{B} = \frac{1}{|B(\beta_i)|} \sum_{x_j \in B(\beta_i)} x_j \quad ..(1)$$

bi represents the i[th] cluster associated with the centroid vi. B(βi) and A(βi) denote the lower bound and the boundary region of cluster βi. The parameter w and w̃ correspond to the relative importance of lower bound and boundary region, and w + w̃ = 1. The main steps of rough c-means are as follows:

1. Assign initial means vi,i=1,2,⋯,c. Choose value for threshold δ.
2. For each object xj, calculate distance dij between itself and the centroid vi of cluster βi.
3. If dij is minimum for1≤i≤c and (dij−dkj)≤δ, then xj ∈ A(βi) and xj ∈ A(βk). Moreover, xj is not part of any lower bound.
4. Otherwise, xj ∈ A sets, xj ∈ A(βi).
5. Compute new centroid as per Equation 1.
6. Repeat steps 2 to 5 until no more new assignments can be made.

## VI. Result of Clustering

Implementation of Rough C-Means shows that if we indicate no of classifier in which data sets should be classified then we can retrieve knowledge from them. The common attributes of elements is likely analogous to the cluster of same property data and can be operated to identify the common property attribute in it. The collection of those attribute leads to the information about the range of element inheriting the same property. Which intern provides the knowledge about that set of data element in that cluster. These results can be extensively used to provide information regarding the set of elements. No of iteration specifies the max no of iteration, but in our implementation we get our cluster in the 7[th] iteration. In this no of data items specifies the no of data sets we are giving as input. Weight Load (WL) specifies the upper and lower bound of the cluster. In our implementation we also need to specify the no of cluster we want . We made & cluster each cluster having a center point which specifies the points around it belong to the cluster. Points in a cluster have a similar characteristic.

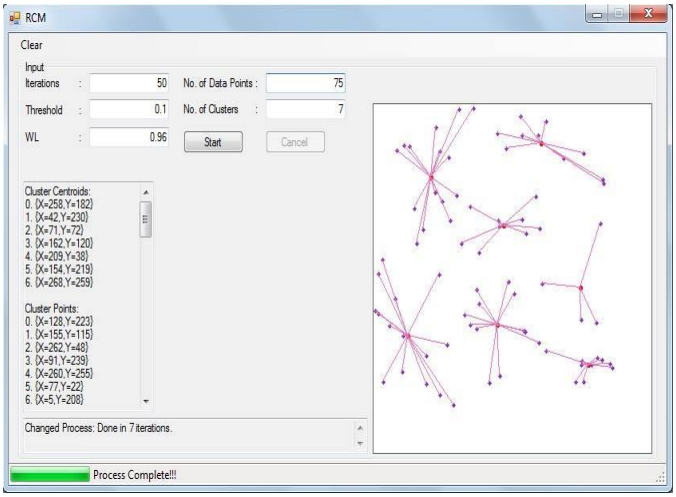


Fig 2. Rough C-Means clustering

## VII. Conclusion and Future Works

Rough C-Means algorithm can be extensively used in mining over the cloud, where the date is huge and is of different of nature. Rough C-Means algorithm can extensively cluster the huge data and the knowledge we get can be used for prediction and decision making. The accuracy of clustering can further be improved with more advance clustering algorithm like RFCM (Rough Fuzzy C-Mean) and can be deployed in a system with more accuracy.

## Acknowledgement

We hereby prolong our sincere gratitude and our hearty regards to our respectful guide Prof. Dr. DHINESH BABU L.D who guided us and helped us in a lot better way to overcome our difficulties in solving this paper and for being a source of inspiration throughout the journey. I am obliged to staff members of VIT, for the valuable information provided by them in their respective fields.

## References

[1] *Fayyad, Usama M., Gregory P., and Padhraic S.,1996. An overview. Advances in knowledge discovery and data mining. CA: AAAI Press . 1-36.*
[2] Hu, Tingting, et al.,2012 "A survey of mass data mining based on cloud-computing.",IEEE,ASID.
[3] Chad West ,Lingras and Pawan.2004. "Interval set clustering of web users with rough k-means." *Journal of Intelligent Information Systems* , 1: 5-16.
[4] Pawlak, Zdzisław.1991. "Rough sets: theoretical aspects of reasoning about data, system theory" .


ISSN: 2278 – 7798 

**Mukundkumar Jha** received the M. Tech degree in software technology from V.I.T University in 2015. Area of interest: Big-Data analytics, Cloud Computing, Data Mining, Machine Learning.

**Deepak Kushwaha** received the M. Tech degree in software technology from V.I.T University in 2015, and is currently working toward the PhD degree in the College of Computer Science,V.I.T university. His research interests include data privacy protection, Data mining, and personalized information retrieval.

**Nawaz Aslam** received the M. Tech degree in software technology from V.I.T University in 2015, and is currently working toward the PhD degree in the College of Computer Science,V.I.T university. His research interests include data privacy protection, Data mining based on cloud computing.

**Amarjeet Singh Yumnam** received the M. Tech degree in software technology from V.I.T University in 2015, Area of Interest: Cloud Computing, Data Mining.



ISSN: 2278 – 7798 