ISSN: 0975-766X
CODEN: IJPTFI
Available Online through
Research Article
www.ijptonline.com

EXTENSION OF ROUGH SET K-MEANS USING PARALLEL MAPREDUCE ALGORITHM

Tanu Maheshwari, M.Nirmala*, Dhinesh Babu L.D, Sharon Moses J
School Of Information Technology and Engineering, VIT University Vellore, Tamil Nadu.
Email: nirmaladhinesh@gmail.com

Received on 25-10-2016 *Accepted on 02-11-2016*

Abstract

Multiple core processors are present on most of the modern computers. For getting greatest benefits of computational power from the core architecture on existing algorithms, we need an advanced design and software. We assert the parallelization of the rough k-means clustering algorithm. In k-means algorithm for clustering by rough set, every cluster is to be formed regarding with the two approximations, lower and upper approximation. For making rough k-means clustering be fine parallelized, we employ parallel mechanism on rough K-means. Therefore, the implementation can be eminently parallel and fault and defect tolerant. The experimental results flaunt bountiful speeduprate of asserted parallel rough k-means clustering method as compared to basic rough k-means algorithm methodology.

Keywords: Clustering, rough set, k-means clustering.

1. Introduction

With the fast advancement of technology and data analysis, the enterprises have a significant amount of data. Deficiency and unclearness of learning arewidespread phenomena in any information system. It is important to manage the fragmented and unclear data in data analysis, order, and concept recognition. The evolution of big data had made people share and store the enormous amount of data [1], [2] The emergence of cloud computing [3],[4] and vital need to manage and retrieve data made clustering one of the needed methodologies of the web 4.0. To attain the objective, numerous hypotheses and methods have been proposed. The rough set hypothesis is an exceptionally successful tool to manage granularity and unclearness in data frameworks. The fundamental goal of clustering will be gathering the information into classes in which the objective is to group different objects into similar kind of sets. Clustering is also an unrestrictedly learning. Clustering can be recognized as partitioning based, hierarchically based, density based, grid based and model based processes.

Information is continuously accumulated and consolidated over an other mixture of fields, at a farcical pace, particularly in theinternet. Rough sets and fuzzy sets [5] acquire two unique aspects of imperfection in identifying data: imperceivably and unclearness. It is a measurable method of information mining and analysis. A rough cluster is a set [6], whose elements can be a part of one or more clusters. The high and low close estimation of the rough set focuses on comparability relations of a set. The low estimation of a rough cluster includes elements that are just having a place with that cluster. The High estimation includes items in the cluster that can also be the members of other sets. Extension of parallel partitioning [7] based on rough set concept process. First This grouping method can be applied to the data in which we for particular data object grouping is not fixed. Joining the advantages of the parallel method with rough set k-means, we can attain higher transforming rate and precision. Density-based clustering methods collocate the object based on the distance metrics. Grid based clustering method carves the object into a finite number of clusters such that the processing time is to be reduced. Parallel Computing and is the key engineering to verify the productivity and precision to understand the gigantic information mining.

2. Background

2.1 Concept:Rough Set

An essential Rough set hypothesis it has been reached out in numerous areas [5][6]. A rough set is a mathematical tool. It is additionally used to recognize the partial or total dependencies in information furthermore wipe out redundancies. It uses the concept of Equivalence relation of classes and close estimation details of classes. It distinguishes data set in Lower and upper close estimation areas.

2.2 Approximation criteria: Lower and Upper

From rough set recognizes the unclearness in the particular data set or area. In rough clustering, every cluster has two typeof approximations, a lower and above approx criteria. Upper set is a superset of lower set. The data parts of the lower rough cluster have an existence surely to the top cluster. The data set which exists in an upper cluster can exist in any other part of the cluster. The common part of lower cluster region and higher cluster region gives the roughness of data objects. Let W be the called as the wholeset; We cannot equal to null and LAR be an equivalence relation for the entire set. For every data object belongs and close estimation criteria for the relation LAR is defined as

Lower Set Criteria:

$$\underline{\boldsymbol{LAR}} = \{\boldsymbol{y} \in \boldsymbol{W} \mid [\boldsymbol{Y}]_{\boldsymbol{LAR}} \subseteq \boldsymbol{Y}\}$$

Moreover,Upper cluster criteria of Y for relation LAR defined as follows:

$$\overline{\boldsymbol{LAR}} = \{\, \boldsymbol{y} \in \boldsymbol{W} \mid [\boldsymbol{Y}]_{\boldsymbol{LAR}} \; \cap \; \boldsymbol{Y} \; \neq \boldsymbol{\Phi}\}$$

Let us consider the example of data set:

In Which we assume Equivalent classes

$\{\{P1, P5\}, \{P2\}, \{P3, P9\}, \{P4, P7\}, \{P6\}, \{P8\}\}$

Let the final resultant Class

$\{P1, P2, P3, P4, P5, P9, P10\}$ $\overline{LAR} = \{\{P1, P5\}, \{P3, P9\}, \{P2\}, \{P4, P7\}\}$

Roughness according to rough sets

$\overline{LAR}$- $\underline{LAR} = \{P4, P7\}$

3. Related Work

Rough set which is a tool that deals with the uncertainty and unclearness. This rough set theory that is applied to specific clustering methods can take one benefit of multicore processors present in that programs implementing its methods may easily run in parallel.

The study of parallel k-means algorithm adopted.

3.1 Basic K-means

The basic K-means is a standout amongst the most popular algorithm for clustering. K-means algorithm divides the set of object samples into specified k clusters and finds the cluster centroid.

1. Arbitrarily k points selection as the cluster centers (initial).
2. Calculate the Euclidean distance for every point taken and accordingly, assign a value to the nearby cluster midpoint.
3. Repeat step 2 for each data point.
4. Compute the average value of the each point in that group.
5. Repeat Steps 2 and four until the data set elements not fixed.
6. Find out the difference between the recent center of the cluster and first center.
7. If changes in observed value do not make a major difference stops the process. Otherwise, recalculate step 2 and 4.

3.2 Adoption of Rough k means clustering:

The core concept is overlapping of data elements into the clusters [8]. When we have an unclear specification for data objects, rough k means the conceptis used.

The rough K-means algorithm:

a)Set k clusters

b) For calculation of centers, identify objects intolower SL(s) and upper SU(s) estimation group.

C)Computation of the centroids of groups from basic k-means needs to be altered to incorporate the impacts of lower and also upper Rough Approximation.

D) Then the modified centroid calculations for rough sets are given by:

IF

$$SL(x) \neq \phi \text{ and } \mathrm{SL(X) - SU(X)} = \emptyset$$

THEN$\frac{\sum_{v \in SL(x)} v_j}{|SL(x)|}$

ELSE IF $SL(x) = \phi \text{ and } \mathrm{SL(X) - SU(X)} \neq \emptyset$

THEN $\frac{\sum_{v \in SU(x)-SL(x)} v_j}{|SU(x)-SL(x)|}$

ELSE

$$W_L \times \frac{\sum_{v \in SL(x)} v_j}{|SL(x)|} + W_U \times \frac{\sum_{v \in SU(x)-SL(x)} v_j}{|SU(x) - SL(x)|}$$

Where$W_L + W_U = 1$ these are corresponding weightage of higher and lower clusters.

e) Calculate the closest center:

$$\mathrm{distance(v1, xi)} - \mathrm{distance(v1, xj)} \leq \text{set threshold}$$

$$v \in SU(x_i) \& \, v \in SU(x_j).$$

f) otherwise v∈ SL(xi), and also v∈ SU (xi).

g) If cluster results same as the previous one then stop the process or continue steps c to d

Some characteristics of rough k-means iterations:

1. At most one data object can be a part of one lower cluster.
2. If specific data set is part of lower cluster, then it is necessary that it should be part of higher group
3. If a data element does not belong to lower cluster, then it must belong to atleast one higher group

3.3 Limitation

Rough K-means clustering algorithm has some limitations. The basic K-means grouping is efficient, but we cannot rely on the time since it might change excessively and the effectiveness can also vary very much and reach a low value when used to process large information [9]. At the point when taking care of the enormous information, the memory of a single hub additionally can be a limitation [10]. As a result, it has not been utilized formerly with extensive information sets.

So it is required to consolidate this conventional calculation with MapReduce skeleton to actualize the parallel calculation.

4. Experiment Proposed

An example contains four subjects is considered and each subject contains two attributes named as anS1 index and S2 index as seen in Table 1.

Assumed value No of the clusters: (k=2)

Table 1: S1 Index and S2 Index

Object	Attribute1(X1) S1 index	Attribute2(Y (Y1) s2 index	Final (result)
Subject P	1	1	c1
Subject Q	2	1	c1
Subject R	5	3	c2
Subject S	3	4	c2

4.1 Rough K-Means

For thevalue of k=2 results in steps involved.

Step 1: Initially assume

Step 2: Initially assume the center values $C_1(1,1)$and

$C_2(2,1)$and a threshold value of 0.5.

Step 3: By the Euclidean distance as in basic k-means:

$$d(i,j)=\sqrt{(|y_{i_1}-y_{j_1}|^2+|y_{i_2}-y_{j_2}|^2+...+|y_{i_p}-y_{j_p}|^2)}$$

Step 4:Applying criteria on higher and the lower set until the iterations result not repeated.

Calculate distance of each point from cluster

Assume cluster centers as

C1(1,1), C2(2,1)

D(P,C1)=0 D(P,C2)=1 D(Q,C1)=1 D(Q,C2)=0 D(R,C1)=3.67

D(C,C2)=2.87

For each object V ε A,B,C,D

I. If D(V,C1)-D(V,C2)<=Max then

- V εSU(C1) and V εSU(C2)
- V does not belong to any lower approximation.
- V εSSL(C) such that C=Min{D(V,C1),D(V,C2)}and V ε U(C) .

4. Result

P εSU(C1) and P εSU(C2) and Q εSL(C2) and Q ε U(C2)

R εSU(C1) and R εSU(C2) and S εSU(C1) and S εSU(C2)

In figures 1, 2 and 3, the results of running the elements can be seen

Figure 1: Projection of elements after Run 1

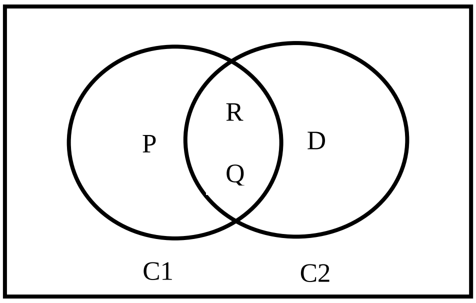


Figure 2: Projection of elements after run 2

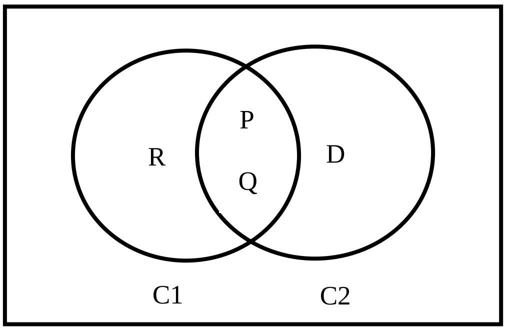
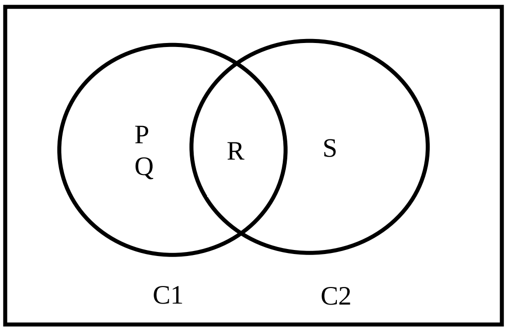


Figure 3:Projection of elements after run 3



5. Proposed Work

The basic rough k-means takes more time in theanalysis the distance between each data sets and recent meancenter for proposed cluster. Both the separation and mean count forms must be rehashed more times, so we plan our calculation to enhance the execution by isolating the work of both of these methods into processes, and after that compute the qualities from every form before running in the following step. Division methodologies of count into methodology are the premise of the parallel method. Figure 5 portrays the entire proposed work.

Improved Parallel rough K-means Algorithm

K-means parallel processing:

calculation is partitioned into the accompanying stages.

Start:Randomly initialize k centers

¨ L1 (0) = L2(0), …, Lk(0)

Step 1:Classify: Assign each point j {1, …m} to nearest center:

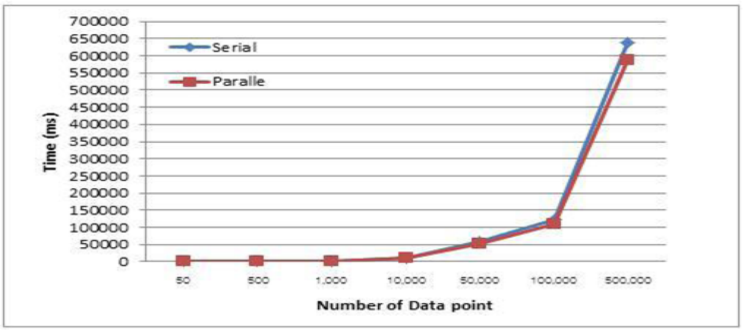
Step 2: Recenter: Li becomes centroid of its point:

Step 3: Equivalent to µiaverage of its points.

C) Identify the difference between the new centroids with the previous centroids in the same cluster. In figure 4 comparison can be seen.

Comparison Graph:

Figure 4. Comparison Set.



Step 4: Merge Function: Select data set{a1,a2,a3,...,an}. At that point partition the entire dataset to sub data clusters such as part1, part2, part3...part n. Then form into <No, Value> lists.

a) Update the clustering centers if needed

b) Rearrange every information into the closest cluster until all the information have been prepared.

C)Sum (or Max) P versions of center sums, point sums, stop criteria.

The map can run in parallel.

c) Output <ci, xj> pair. Moreover, Li is the center of the

cluster Ci.

<No, Value> -->input into map procedure.

Step 5: Calculate High Close estimation of each member in K clusters

Step 6:Recalculate new centroid N'.

If difference(N, N') Sequential code use mean directly not SumsThen set N to be N' and go back to step 2

Step 7:Assume the threshold value $\mathrm{distance}(v1, xi) - \mathrm{distance}(v1, xj) \leq \text{set threshold}$

Step 8:**Reducerfunction**

a) From the output of Map stage. Collect data records.

Moreover, then output the k clusters and the data points.

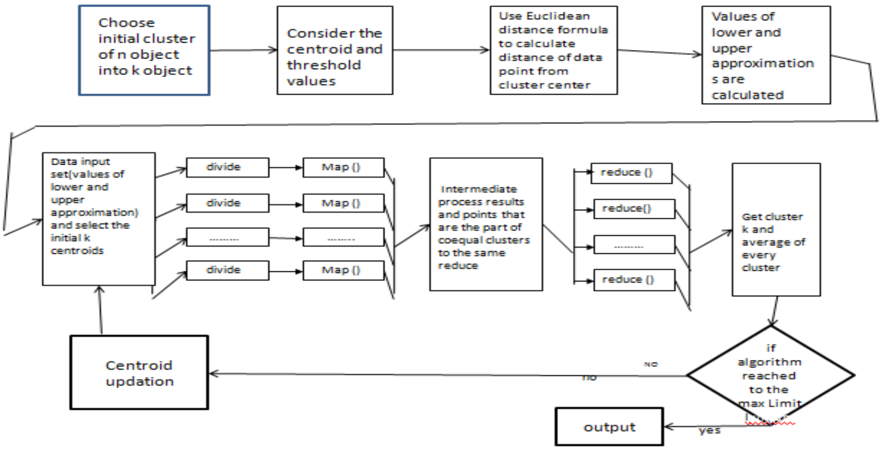
b) Again identify the average of each group which is chosen as the new cluster center.

Step 9: For each Run, Send results back to parent process

a)Calculation has arrived at the most extreme limit,

Stop the process

Figure 5. Final Proposed Rough set K-me and Parallel clustering diagram.


Choose initial cluster of n object into k object
Consider the centroid and threshold values
Use Euclidean distance formula to calculate distance of data point from cluster center
Values of lower and upper approximations are calculated
Data input set(values of lower and upper approximation) and select the initial k centroids
divide
divide
....
divide
Map ()
Map ()
....
Map ()
Intermediate process results and points that are the part of coequal clusters to the same reduce
reduce ()
reduce()
....
reduce ()
Get cluster k and average of every cluster
if algorithm reached to the max Limit
NO
no
yes
Centroid updation
output

6. Conclusion

Rough k-means clustering is a clustering technique that is ascertained on parallel clustering to help in approximate clustering with data with ambivalent membership states. The clustering estimations are the lower and upper approximation. Rough Clustering is augmenting popularity for potential use in Web usage
In this paper, we assert the algorithm and pursuit the rough k-means clustering algorithm in theparallel style rough set.

References

1. Krishna, P. V. Honey bee behavior inspired load balancing of tasks in cloud computing environments. Applied Soft Computing.2013;13(5): 2292-2303.
2. Raj E.D, Babu L.D, Area E, Nirmala M, Krishna,P.V. Forecasting the Trends in Cloud Computing and its Impact on Future IT Business. Green Technology Applications for Enterprise and Academic Innovation; 2014.p.14.
3. Babu L. D, & Krishna P. V. An execution environment oriented approach for scheduling dependent tasks of cloud computing workflows. International Journal of Cloud Computing. 2014; 3(2): 209-224.
4. DhineshBabu L. D, Gunasekaran A, Krishna, P. V. A decision-based pre-emptive fair scheduling strategy to process cloud computing work-flows for sustainable enterprise management. International Journal of Business Information Systems. 2014; 16(4): 409-430.
5. PawlakZ. Rough set theory and its applications to data analysis. Cybernetics & Systems. 1998;29(7):661-688.
6. Pawlak Z. Rough sets - Theoretical aspects of reasoning about data, Dordrecht:Kluwer Academic Publishers.1991;1(1):68-162.
7. Devine K, Boman E, Heaphy R, Hendrickson B, Vaughan C. Zoltan data management services for parallel dynamic applications. Computing in Science & Engineering.2002; 4(2): 90-96.
8. Jensen R, Shen Q, Semantics-preserving dimensionality reduction: rough andfuzzy-rough-based approaches. IEEE Transactions on Knowledge and Data Engineering.2004;16 (12):1457-71
9. William Z, Fei-Yue Wang, F on Three Types of Covering-Based Rough Sets. IEEE Transactions on Knowledge and data engineering. 2007;19(8): 1131-1144.
10. Zhong, N, Dong J, Ohsuga, S. Using rough sets with heuristics for feature selection. Journal of intelligent information systems.2001;16(3):199-214.