ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 6, June 2015

An Enhance Approach to Improve CURE Clustering Using Appropriate Linkage Function for Datasets

Ms Komalben N. Makadiya, Prof. Maulik V. Dhamecha

M.Tech Student, Department of Computer Engineering, RK University, Rajkot, India

Assistant Professor, Dept. of C.E., RK University, Rajkot, India

ABSTRACT: Clustering is an unsupervised learning process. Clustering algorithms classify data points into meaningful groups based on similarity. There are various clustering methods. CURE is an Agglomerative Hierarchical Clustering algorithm. CURE is Clustering Using Representatives. CURE is finding the clusters from a large database that is tougher to outliers and classifies clusters having non-spherical forms and varied differences in size. CURE works a grouping of data reduction & data collection by using random sampling and dividing. So, CURE is handling to large databases. In this paper, CURE and Improved CURE were examined and analysed based on the Linkage Criteria and Distance measure. This proposed algorithm is an applied and evaluated using a clustering device WEKA.

KEYWORDS: Data Mining, Clustering Algorithms, CURE Algorithm, WEKA tools

I. INTRODUCTION

Data Mining is a method that usages a variation of data examination tools to determine patterns and relations in data. Data Mining denotes to extracting or mining knowledge from huge amounts of data [1]. Clustering is discuss to the method of grouping objects into within the same class are similar in a certain logic and objects from different classes are dissimilar [4]. Clustering algorithms can be useful in many areas including Biology, Earthquake Studies, Insurance, Marketing, Libraries and City-planning. Clustering techniques are focused on scalability, effectiveness, difficult shapes and categories of data, high dimensional data, numerical and categorical data in huge databases [7].

Hierarchical clustering creates a hierarchy of clusters which may be denoted in a tree of structure is called a dendrogram [1]. Clustering output can be created by many tools like WEKA, XLMiner, Orange, KNIME and Tanagra but here we select WEKA (Waikato Environment for Knowledge Analysis) for implementing our hierarchical clustering algorithm [3]. WEKA toolkit is a generally used for machine learning and data mining. It was firstly established at the University of Waikato in New Zealand [16]. Our aim to display the evaluation of the clustering algorithms on WEKA and finding out which algorithm will be best appropriate [3].This paper determines improvements of CURE Clustering algorithm.

II. RELATED WORK

A Comparative Study of Various Clustering Algorithms in Data Mining

Mauly Srivastava, Atul Kumar Diswar, Manish Verma, Nidhi Gupta [10].In this paper, There are six types of Clustering techniques-K-means Clustering, Hierarchical Clustering, DB Scan clustering, Density Based Clustering, Optics, EM Algorithm. These clustering techniques are executed and analysed using a clustering device WEKA.

Comparision the various clustering algorithms of weka tool

Aman Bajpai, Narendra Sharma, Mr. Ratnesh Litoriya [16].We are revising the various clustering algorithms. Clustering is the task of conveying a set of objects into groups called clusters so that the objects in the similar cluster are more similar to each other than to those in other than to those in other clusters. Our main goal is to show the

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 6, June 2015

comparisons of the various clustering algorithms on WEKA. Then find out which algorithm will be most proper for the users.

Dynamic Clustering of High Speed Data Streams

J.Chandrika, Dr. K.R. Ananda Kumar[4]. In this paper, we offered research activities and methods that have developed in the field of data stream mining. The research in this field is mainly done in modelling, mining and query processing data streams. The application of experimentally developed approaches, techniques and methods is determined by the problem domain and the properties of problem domain.

CURE: AN EFFICIENT CLUSTERING ALGORITHM FOR LARGE DATABASES

Sudipto Guha, RajeevRastogi and Kyuseok Shim [2].In this paper, We talked problems through customary Clustering algorithms. They are approval clusters with spherical shapes and similar sizes. CURE uses more than one representative point for every cluster. CURE modifies well to the geometry of clusters taking on-spherical shapes and wide differences in size. CURE works a combination of random sampling and partitioning. So, it handles large datasets well**.**

A Survey of Clustering Algorithms for Big Data: Taxonomy & Empirical Analysis

A.Fahad, N.Alshatri, Z.Tari, A.Alamri, I.Khalil, A.Zomaya, S.Foufou and A.Bouras [13].This review provided a complete study of the clustering algorithms planned in the literature. We suggested a classifying framework to order a number of clustering algorithms. The Classified framework is established from a theoretical viewpoint that would spontaneously mention the most appropriate algorithm to network experts while hiding all practical details unrelated to an application.

III. CLUSTERING ALGORITHMS

Clustering Algorithms can be divided in various categories. This are Partitioning based, Grid based, Model based, Hierarchical based and Density based [13]. Hierarchical Clustering are further separated in Agglomerative and Divisive. In Agglomerative Clustering which one starts at the leaves and sequentially combines cluster together. In Divisive Clustering which one starts at the root and recursively separations of the clusters, Hierarchical technique tells to the fact that once a step is executed, this cannot be undone [10].

A. CURE Clustering Algorithm

CURE works a new hierarchical clustering algorithm. CURE varies well to the geometry of non-spherical shapes. CURE is less sensitive to outliers [4]. It expenditures a fixed number of points as representatives (partition).It implements a middle ground between representative-object based and centroid-based approaches[3]. A constant number c of well distributed ideas in clusters are selected and then shrunk to the center of the cluster by identified portion alpha. CURE agrees more than one representative point per cluster. The clusters with the neighboring pair of symbolic points are combined at each step. It breaks when there are only k clusters left [2].CURE can recognize arbitrarily shaped clusters. This algorithm usage linear space in the input size n and takes a worst-case time complexity of O (n2 log n). For minor dimensions, the complexity can be reducing to O (n2) [6].

Steps for CURE Algorithm

a) Draw random sample

We decrease the size of the input to CURE in order to perform huge data sets. CURE is applied to an arbitrary sampling drawn after the dataset rather than the full data set. It can progress the value of clustering because it takes the essential effect of filtering outliers [2].

b) Partitioning for Speedup

The partings among clusters are reduces and clusters are developed less closely filled. Partition that sample according to dimensions. Divide the selected space into p dividers, each of size n/p [2].

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 6, June 2015

c) Partially cluster partition
Each division incomplete clustering while waiting for the last number of clusters in every divider decreasing n/ (p q). (q>1) [2]

d) Eliminate Outliers
Firstly, unselected sampling cleans out a popular of the outliers. Before, the outstanding outliers are spread completely the sample space and gets advance isolated. Originally, each point is a discrete cluster and continues by combining nearby points first. They are larger distances from other points attend to join another points fewer and developing very slowly are recognized and removed as outliers [2].

e) Labeling Data on Disk
Element is an arbitrarily chosen by sample. It needs to allot the correct cluster labels to the left over data points. From each one data point works a portion of arbitrarily designated representative points for each of the ultimate k clusters. Every cluster with some points alternatively a single centroid supports CURE to properly allot the data points while clusters are non-circular or non-regular.

Merits & Demerits of CURE:
Merits
CURE can detect cluster with non-spherical shape and widespread alteration in size using a group of representative points of each cluster.
CURE can also be good in execution time in presence of large database using random sampling and partitioning methods.
CURE works well when dataset contain outliers. They are identified and removed.

Demerits:
Consider only one point as representative of a cluster.

B. Proposed algorithm

Our proposed algorithm that reduce the time using Linkage functions and distance measure method that gives good result as compare to CURE clustering algorithm.

Steps of Improved CURE

1) Take numerical data set as an input
2) Use Chebyshev distance for linking weights
3) Construct the tree
4) Now partitioning clusters
5) Apply Clustering using k-Nearest neighbor joining linking.
6) After Partioning merge the partitions.
7) Apply label on outputs.

IV. CLUSTER DISSIMILARITY

We have to choose which clusters should be joined (for agglomerative), or where a cluster should be split (for divisive), a measure of difference between groups of observations is necessary. In hierarchical clustering, this is completed by usage of an appropriate metric and a linkage criterion.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 6, June 2015

A. Distance Metrics:

It is a measure of distance amongst pairs of observations.

1) Euclidean Distance

It calculates the root of square modifications between coordinates of two points. If p = (p1, p_2,…p_n) and **q** = (q_1, q_2,... q_n) is denoted as[10]:

N dimensional Euclidean distance

. $d(p,q)=\sqrt{(p_1-q_1)^2+(p_2-q_2)^2+(p_3-q_3)^2+\ldots+(p_n-q_n)^2}$

- $$d(p,q)=\sqrt{\sum_{i=1}^{n}(p_i-q_i)^2}$$

2) Manhattan distance or City block distance

It denotes distance between two points in a city road grid. It calculates the absolute differences among coordinates of two objects. The distance between two points measured along axes at right angles. In a plane with A_1 at (p_1, q_1) and B_1 at (p_2, q_2) is denoted as [15]:

N-dimensional Manhattan distance

$$d(p,q)=\sum_{i=1}^{n}|p_i-q_i|$$

3) Minkowski distance

It is a generalization that joins Euclidean distance, Manhattan distance, and Chebyshev distance. The Minkowski distance of order c between two points **p** = (p_1, p_2,.., p_i) and **q** = (q_1, q_2,..., q_i) is denoted[15]:

N-dimensional Minkowski distance

$$d(p,q)=\left(\sum_{i=1}^{n}|p_i-q_i|^c\right)^{\frac{1}{c}}$$

4) Chebyshev distance

It is also known as Extreme value distance. It estimates absolute magnitude of the changes between coordinates of two points. It is also known as Chessboard distance because in the game of chess the least number of changes required by a king to go from one square on a chessboard to another equals [14].

$$d(p,q)=max_i|p_i-q_i|$$

B. Linkage Methods

It identifies the dissimilarity of sets as a function of the pairwise distances of observations in the sets.

1) Single Linkage:

The distance between two clusters is smallest distance between an observation in one cluster and other cluster. [15]

$$D(X,Y)=\min_{x\in X,y\in Y}d(x,y)$$

2) Average Linkage:

The distance among two clusters is mean distance between an observation in one cluster and other cluster. [15]

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 6, June 2015

$$D(X,Y) = \frac{1}{|A| \cdot |B|} \sum_{x \epsilon A} \sum_{x \epsilon B} d(x,y)$$

3) Complete Linkage:

The distance among two clusters is extreme distance an opinion in one cluster and other cluster [15].

$$D(X,Y) = \max_{x \in X, y \in Y} d(x,y)$$

4) Centroid Linkage:

The distance between two clusters contains one point centre of cluster to other cluster centre points (centroids or means).

V. EXPERIMENTAL SETUP AND RESULTS

The performance of future algorithm is verified across four data sets and all of which covers only numeric attributes and class attributes. Algorithms require their executed source code in the WEKA 3.7.10 Version. They have approved out in order to measure the performance parameters of the algorithms over the datasets. Our experiments were performed on a 64-bit Windows-based system with Intel core (i3), 1.80 GHz processor machine with 4 G bytes of RAM. The simulation has been made on the platform of the WEKA APIs and Eclipse IDE.

We evaluate the performance of clustering process by building a model using training data set. The results are derived as time taken to output clusters and number of data points. The information about the data sets and Execution time are tabulated in following Table.

TABLE I. TIME REQUIRED TO BUILD MODEL

Dataset	Number of Points	Attributes	CURE	Improved CURE
Pima-diabetes	1007	9	4340	1240
Wine	1450	12	8150	3350
Ionosphere	2279	35	23620	15390
Image Segment	2338	20	13450	13380

The execution time for each algorithm was also evaluated and the results were matched with one another, both CURE and proposed CURE clustering methods were observed to study based on the distance between the several input data points. The clusters were designed according to the distance amongst data points and clusters centers were calculated for each cluster. The data points in each cluster were shown by different colors and the execution time was calculated in milliseconds.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 6, June 2015

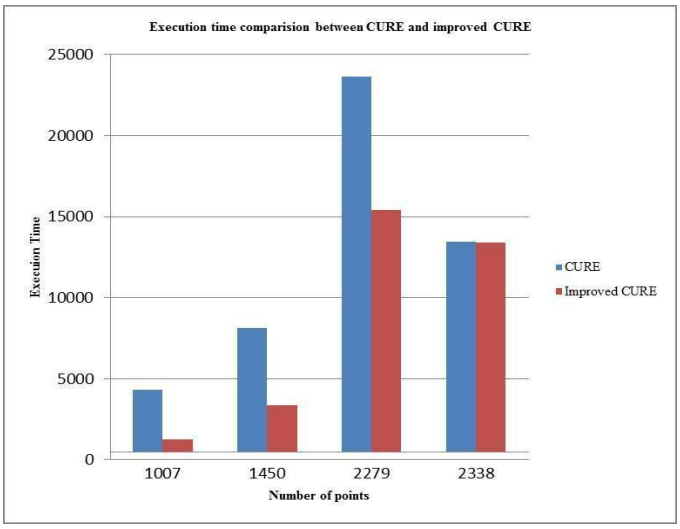

Fig.1. Time taken to build model

VI. CONCLUSION AND FURTHER WORK

We have proposed a modification in the CURE algorithm and the experiments prove that with this improve performance increase, by changing the distance similarity measure. The performance of proposed algorithm is verified across four real world datasets and the results have proven the efficiency of the proposed algorithms. So, over all it will be good algorithm for clustering with aspect of handling large datasets and time parameters. From the result it is proven that our approach is better in data handling as well as in time. Further work can be done as to compare the results obtained by the hierarchical clustering algorithms proposed in this paper with other clustering approach.

VII. ACKNOWLEDGMENT

I would like to thanks Prof. Maulik V. Dhamecha for his help and the Department of Computer Engineering of RK University, Rajkot.

REFERENCES

1. H. Han and Kamber, "Data Mining : Concepts and Techniques", Morgan Kaufmann Publishers, 2001.
2. Sudipto Guha, Rajeev Rastogi and Kyuseok Shim, "CURE: An Efficient Clustering Algorithm For Large Databases", Elsevier Science Ltd, Information System Vol. 26,No. 1,pp. 35-58,2001
3. Yogita Rani, Manju, Harish Rohil, " Comparative Analysis of BIRCH and CURE Hierarchical Clustering Algorithm using WEKA 3.6.9", The Standard International Journals (The SIJ), 2014, ISSN: 2312 – 2381
4. J. Chandrika1, Dr. K.R. Ananda Kumar, "Dynamic Clustering Of High Speed Data Streams", IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 2, No 1, March 2012.
5. Sandra Sagaya Mary. D. A, Tamil Selvi. R," A Study of K-Means and Cure Clustering Algorithms", International Journal of Engineering Research & Technology (IJERT), Vol. 3 Issue 2, February – 2014.
6. Muhammad Husnain Zafar and Muhammad Ilyas," A Clustering Based Study of Classification Algorithms", International Journal of Database Theory and Application (IJDTA), Vol.8, No.1 (2015), pp.11-22.
7. Komal N. Makadia, Prof. Maulik V. Dhamecha," A SURVEY ON ENHANCING AGGLOMERATIVE HIERARCHICAL TECHNIQUES", International Journal of Advance Engineering and Research Development Volume 1, Issue 11, November -2014.
8. Amandeep Kaur Mann, Navneet Kaur," Survey Paper on Clustering Techniques", international Journal of Science, Engineering and Technology Research (IJSETR) Volume 2, Issue 4, April 2013.
9. Amita Verma, Ashwani kumar," Performance Enhancement of K-Means Clustering Algorithms for High Dimensional Data sets", International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE), Volume 4, Issue 1, January 2014.
10. Manish Verma, Mauly Srivastava, Neha Chack, Atul Kumar Diswar, Nidhi Gupta ," A Comparative Study of Various Clustering Algorithms in Data Mining", International Journal of Engineering Research and Applications (IJERA) ,Vol. 2, Issue 3, May-Jun 2012, pp.1379-1384.
11. Pavel Berkhin, A Survey of Clustering Data Mining Techniques, Yahoo!, Inc. pberkhin@yahoo-inc.com
12. UCI Machine Learning Repository/ http://archive.ics.uci.edu/ml/
13. A. Fahad, N. Alshatri, Z. Tariz, A. Alamri, I. Khalil A. Zomaya,, S. Foufou, and A. Bouras, A Survey of Clustering Algorithms for Large Data: Taxonomy & Empirical Analysisl, IEEE Transactions on Emerging Topics in Computing, Volume: PP 1-12,12 June 2014,ISSN :2168-6750
14. Grabusts,Peter, "Distance Metrics Selection Validity in Cluster Analysis",Computer Science(1407-7493),2011.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 6, June 2015

15. Wikipedia http://en.wikipedia.org.
16. Narenda Sharma, Aman Bajpai, Mr.Ratnesh Litoriya,"Comparison the various clustering algorithms of weka Tools", International Journal of Emerging Technology and Advanced Engineering (IJETAE), Volume 2 Issue 5 May 2012.

BIOGRAPHY

Ms. Komalben N. Makadiya is a M.tech Student in the Computer Engineering Department, R.K. University, Rajkot, India. She received Bachelor of Information Technology (BE (IT)) degree in 2013 from V.V.P. Engineering College, Rajkot, India.

Prof. Maulik V. Dhamecha is an Assistant Professor in the Computer Engineering Department, R.K. University, Rajkot, India.