INTERNATIONAL JOURNAL FOR INNOVATIVE RESEARCH IN MULTIDISCIPLINARY FIELD ISSN: 2455-0620 Volume - 6, Issue - 6, June – 2020
Monthly, Peer-Reviewed, Refereed, Indexed Journal with IC Value: 86.87 Impact Factor: 6.497
Received Date: 03/06/2020 Acceptance Date: 18/06/2020 Publication Date: 30/06/2020

Cure Clustering Algorithm and Its Various Implementations

[1]Nikita Kumble, [2]Vandan Tewari
[1]Student, [2]Associate Professor
Department of Computer Engineering,
Shri Govindram Seksaria Institute of Technology & Science, Indore, India
Email - [1]nikitakumble@gmail.com, [2] vandantewari@gmail.com

Abstract: *Clustering is an unsupervised learning task of data mining which allows forming clusters of the points which have similarity among them more than the points of other clusters. The traditional algorithms favour the clusters of spherical shapes. Also, they are very sensitive to outliers. Unlike them, the Cure clustering algorithm identifies the clusters of non-spherical shapes and wide variance and can cluster very large size of the dataset. The algorithm selects well-scattered points from the dataset as representative points and shrunk them towards the centre of the cluster to find the clusters. Using multiple points instead of one centroid, it can identify the non-spherical shape clusters and the shrinking of points towards the centre allows it to reduce the impact of outliers. Due to these capabilities, it overcomes the drawbacks of traditional algorithms. In this paper, we are discussing the various implementations of cure clustering algorithms in various domains including outlier detection, non-spherical shaped dataset clustering and to speed up the execution of algorithm and presenting their appropriateness for applications.*

Key Words: *Clustering, Large dataset, Non-spherical Shape, Outlier Elimination, Partition, Sampling.*

1. INTRODUCTION:

Nowadays, a huge amount of data is captured rapidly from organizations and repositories are increasing continuously in size. Analysis of such huge data is computationally inefficient. Data mining allows discovering the knowledge from this data which is useful, to utilize it for future computations. Classification[1] is a supervised machine learning task which contains previously categorized training dataset. Whereas clustering is unsupervised data analysis where there is no training set and characteristics of similarity of data are not known. In clustering, the extraction of hidden patterns in the data set with no labelled response is achieved. It is widely used in various applications such as market segmentation, bioinformatics, voice mining, image segmentation, spatial data analysis, patterns recognition, text mining etc. In cluster analysis[2], smaller the inter-cluster distance and larger the intra-cluster distance forms more distinct clusters. In a cluster, the data points have more similarity between each other than the points of different clusters. Clustering is important in various ways as clusters show the internal structure of the data. It is an important part of the KDD process. It is useful when partitioning of the data is the goal. Clustering prepares the data for other Artificial Intelligence techniques and can be used alone over dataset to understand the distribution, to observe each cluster's features, and after which focuses on the analysis of the clusters. Also, it is used for various data mining techniques like classification and features algorithm as a pre-processing step, and it can be used for further correlation analysis. The traditional clustering algorithms have the drawback of not performing appropriately in datasets having non-spherical clusters, varying densities and in handling outliers as they use either centroid based or all point based approach. Centroid[3] represents the centre of the cluster. Centroid are used to decide the similarity between two clusters. But for a non-spherical dataset, centroids can't identify the clusters appropriately. Instead of a single centroid, in all-point based clustering multiple scattered points are used to represent the clusters. In all-point based clustering, to represent the cluster, multiple scattered points are used. It enables it to identify the clusters of arbitrary shaped and any size but much sensitive to noise and outliers and smaller variation in the position of data points. To overcome these drawbacks, instead of centroid based or all point based, Cure uses a compromise scheme between these two methods for clustering.

2. CURE ALGORITHM:

CURE (Clustering using REpresentatives) introduced by Guha S.[4], identifies complex-shaped clusters. It is robust to outliers than other algorithms. Cure follows neither centroid based approach nor an all-point based approach. It uses representative points. The representative points shrunk towards the centroid, which dampens the adverse effect of outliers. Fig.1 shows the flow of the Cure algorithm.

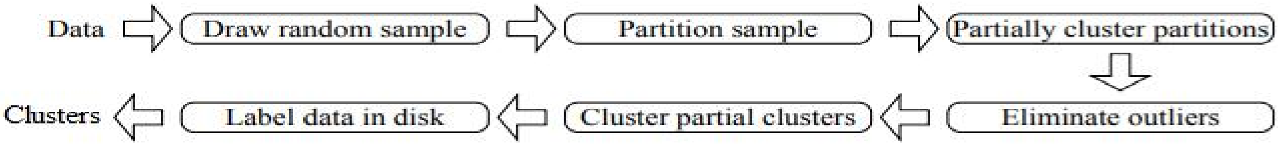

Figure 1. Flow diagram of Cure Clustering Algorithm

2.1. THE PROCEDURE OF CURE ALGORITHM:

Various stages of the Cure algorithm are described as follows:

- Draw a random sample *s*: to handle large dataset, random sampling is done. It reduces the execution time and resolves the problem of short main memory. This filters the outliers which improve the quality of the cluster.
- Partition the sample into *p* partitions: the size of each partition is *s/*p.
- For each partition, partially clustering is done until the clusters in each partition are equal to s/pq.
- Eliminate the outliers: The outliers form very small clusters, so they are easy to observe. So it is easy to identify and eliminate those clusters.
- Cluster the partial clusters to form the required number of clusters.
- Label data on disk: The remaining data points which were excluded from the random sample are assigned to the clusters having representative point closest to them.

For handling the dataset of large size, Cure uses two techniques: sampling and partition. A random sampling of size *s* is used to overcome the problem of shortage of main memory. Instead of the whole dataset, a sample of the dataset of size n which can get fit into the memory is selected for partial clustering. Partition of the data sample is done for faster execution of the algorithm, where the sample space is partitioned into *p* partitions, where the size of each partition is s/p. The algorithm uses two data structure: k-d tree and heap. The representative points are stored in the k-d tree. The heap stores the data points, one from each cluster.

3. RELATED WORK:

3.1 WORK RELATED TO A GOOD SET OF REPRESENTATIVE POINTS TO FIND GOOD CLUSTERS:

In[5], Y. Qian, Q. Shi and Q Wang proposed a new shrinking scheme for the representative points. In their work, instead of shrinking towards the centroid, the direction and distance of shrinking are determined by the density value of representative points. The new shrinking scheme allows to find out appropriate representative points from the random chosen scattered points. For each scattered point, the density value is calculated. The scattered point with maximum density value set as a representative point. This works the same as CURE algorithm but it uses the density value of the scattered points for clustering. T.Aslanidis and D.Souliou[6] uses a new technique to calculate the representative points. This allows overcoming the problem of identifying non-convex shape clusters. They start with each point as a single cluster and merge the two clusters which are closest to each other than others. As two clusters are merged, the points in the new cluster increases. On exceeding the number of points in the cluster, the reduce procedure divides the cluster into zones and selects the point with minimum and maximum value for each zone. It reduces the number of points to the number of representatives in the new merged cluster. To find the two closet clusters, the average value of the closest ndp is used instead of the distance between two closest points.

Zhao Yan[7], proposed an improved algorithm for dynamic analysis of information to identify business competitors. In this work, instead of representative points, they use the identifying element for clustering. The identifying elements are obtained from the features of page content. The weight is the value of identifying element. Various attributes of an object are assigned a weight according to their importance to obtain the major and minor elements. The dynamic adjustment of weight value helps to increase the flexibility of the clustering. They evaluate the identifying element of each competitor to identify weather the competitor is weak or mighty. For clustering, the major element value of competitors is compared to the centre point of the class. If the ratio between them is larger than the threshold, they belong to the same class. After clustering, the merging of small clusters is done by calculating the degree of correlation between the two classes to acquire the required number of clusters.

3.2 WORK RELATED TO SPEED UP THE EXECUTION FOR BIG DATA:

P. Lathiya and R Rani[8] proposed Cure using Apache Hadoop to reduce the execution time by executing it in parallel systems over a distributed environment. The dataset is stored in the HDFS. The input dataset is given in the form of (key, value) pair to multiple Mapper. These pairs are then processed in parallel for the generation of the intermediate result sets which are shuffled and feed to Reducer. The reducer performs the Cure clustering to generate sub-cluster. This sub-clusters are combined until it reduces to the specific number of clusters.

In[9], S.Xiufeng and C.Wei proposed large scale text information clustering using improved Cure. The characteristics of face specified disposal object in text and data of huge size. Cure allows the effective partitioning of the text data for clustering. To overcome the problem of space shortage of main memory by using the traditional algorithm, Cure is used for partitioning the dataset. It allows the huge network to be partitioned according to the resources. To avoid big data problem, for each document, it distils the characters in big document combination. The documents are clustered by calculating the similarity between the document. Cure improves the text clustering by enhancing the precision value.

3.3 WORK FOCUSED ON DETECTION OF OUTLIERS FOR GOOD CLUSTERS:

To detect the network traffic anomaly(DDoS), M Laksono, Y Purwanto and A Novianty[10] propose a new outlier removal clustering(ORC) which uses the Cure clustering algorithm to determine the outlier. The threshold value is used to determine the number of outliers from the dataset. Distortion, which is the ratio of points at a minimum distance and maximum distance respectively from a representative point, is used to identify the outliers. Saravanan[11] uses Cure clustering for video data retrieval. After performing the initial method of data retrieval of video segmentation, image feature extraction and duplicate frame extraction, clustering is done over the image feature value to extract the multimedia data. Random grouping of data points is done after which it generates desired clusters based on the quality of the image. The algorithm scans the image first and identifies the data points in the image. The healthy spotted pixels from the data points are reduced towards the midpoints. This works well for all type of image and video files. The outlier detection quality of cure eliminates outliers to form a good image cluster. The table I depicts a brief comparative study of various implementations of cure algorithm in which a study based on different parameters like used approach, advantages and limitations of particular work has been discussed.

Table: 1 A comparison table for various implementations of Cure Algorithms

S. no	References	Approach	Description	Parameters	Advantages	Limitations
1	CURE algorithm, 1998[4]	Instead of the whole dataset, uses some representative points for clustering non-spherical shaped data which is not centroid based.	Representative points are chosen randomly which shrunk towards their nearest point to form clusters.	Shrinking factor, Threshold	Better efficiency than traditional algorithms. No impact of outliers Identify non-spherical shapes clusters.	Representative points shrunk towards the centroid, hidden assumption of the dataset to be spherical. Sampling does not consider entire dataset, there for the input may not have information about certain clusters.
2	Improved CURE Algorithm for Large-scale Data,2011[9]	Effective data partitioning and hot point finding enable the text clustering efficient.	The web data is partitioned and then each document is a class. The similarity between the two classes is computed to merge the class.	Cosine Distance, TFIDF	Enhances the precision of typical text clustering algorithms.	No threshold for similarity measure. All outlier classes(less than three documents) are stored for the future, but no procedure to handle it.
3	Improved CURE using Hadoop and MapReduce, 2016[8]	To reduce the execution time, the clustering is done in parallel	Data is feed into Mapper using map function in (key, value) form to generate partial clusters. All partial clusters combined in reducer to form final clusters	Euclidean distance, Shrinking factor, Reducing factor	Consumes less time and space Higher efficiency	The dataset must be converted into (key, value) form and stored into HDFS. Focuses on time efficiency only, storage is not considered.
4	CURE-NS 2002[5]	New shrinking scheme to find out a good set of representative points for clustering the dataset.	For all well-scattered points, the points with low density value are shrunk towards the point of having high density value. The points with high density value are stored in the reference set as representative points for clustering.	Density distribution, Shrinking factor	Insensitive to noise and outlier Lower computational time Adaptive to different shapes	If outliers are chosen as the scattered point initially it will be a representative point in the reference set. Assumption that outlier has low density but no threshold for density value, it could not work for clusters with sparse density.

5	DDoS Detection with Outlier Removal Clustering, 2015[10]	Due to efficient outlier detection, Cure is used to performing DDoS detection mechanism.	The threshold is used to determine the count of outliers to be eliminated. Distortion is used to find out the outlier point which causes distortion less than the threshold.	Distortion Threshold Detection rate False Positive Rate	Detects outliers without affecting the detection system capabilities	No specific method to determine the threshold values. The point at maximum distance from a representative point is considered as an outlier, which is not applicable for the non-spherical shaped cluster.
6	CUZ, 2008[6]	Divides the dataset into zones to compute representative points in each zone.	Merge procedure is used to merge the two closest clusters. The reduce function reduces the number of points in a merged cluster to the required number of representative points.	Euclidean distance, ndp(no of distance point)	Identify the clusters of non-convex shapes, Easily adapted for multidimensional space	Distance calculation after each merge operation. The number of representative points in a zone is restricted to two which could be sometimes misleading.

4. RESULT AND DISCUSSION

In this paper, we provide a brief survey of the various implementations of the Cure clustering algorithm. The Cure has various ability over traditional algorithms because of which it is more efficient. The outlier elimination of Cure allows it to detect outliers to form the good quality of clusters. It can identify non-spherical shaped clusters and can handle large-size datasets. We also review the comparative study of these algorithms which allows us to see various applications of cure. This paper provides a snapshot of various implementations of Cure which is very useful to get better performance and better efficiency.

REFERENCES:

1. Bindra, K. and Mishra, A. (2017). A detailed study of clustering algorithms. 6th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, 2017, pp. 371-376. doi:10.1109/ICRITO.2017.8342454.
2. Deshmukh, M. A. and Gulhane, R. A. (2016). Importance of Clustering in Data Mining. International Journal of Scientific & Engineering Research, Volume 7, Issue 2, February-2016 ISSN 2229-5518.
3. Mary, D. A. S. S. and Selvi, R. T. (2014). A Study of K-Means and CURE Clustering Algorithms. INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH TECHNOLOGY (IJERT) Volume 03, Issue 02 (February 2014).
4. Guha, S. Rastogi, R. and Shim, K. (1998). CURE: An Efficient Clustering Algorithm for Large Databases. In: Proceedings of the 1998 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'98). Seattle, Washington, 1998 ISBN:0-89791-9955 doi:10.1145/276304.276312
5. Qian, Y. T. Shi, Q. S. and Wang, Q. (2002). CURE-NS: a hierarchical clustering algorithm with new shrinking scheme, Proceedings. International Conference on Machine Learning and Cybernetics, Beijing, China, 2002, pp. 895-899 vol.2. doi: 10.1109/ICMLC.2002.1174512.
6. Aslanidis, T. Souliou, D. and Polykrati, K. (2008). CUZ: An Improved Clustering Algorithm, *IEEE 8th International Conference on Computer and Information Technology Workshops*, Sydney, QLD, 2008, pp. 43-48. doi: 10.1109/CIT.2008.Workshops.118.
7. Yan, Z. (2010). Research of an improved cure algorithm used in enterprise competitive intelligence to dynamic identify analysis, *IEEE Youth Conference on Information, Computing and Telecommunications*, Beijing, 2010, pp. 299-302. doi: 10.1109/YCICT.2010.5713104
8. Lathiya, P. and Rani, R. (2016). Improved CURE clustering for big data using Hadoop and Mapreduce, International Conference on Inventive Computation Technologies (ICICT), Coimbatore, 2016, pp. 1-5. doi: 10.1109/INVENTIVE.2016.7830238.
9. Xiufeng, S. and Wei, C. (2011). Improved CURE algorithm and application of clustering for large-scale data, IEEE International Symposium on IT in Medicine and Education, Cuangzhou, 2011, pp. 305-308. doi: 10.1109/ITiME.2011.6130839
10. Laksono, M. A. T. Purwanto, Y. and Novianty, A. (2015). DDoS detection using CURE clustering algorithm with outlier removal clustering for handling outliers, International Conference on Control, Electronics,

Renewable Energy and Communications (ICCEREC), Bandung, 2015, pp. 12-18. doi: 10.1109/ICCEREC.2015.7337029.

11. Saravanan, D. (2016). CURE clustering technique suitable for video data retrieval, *IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*, Chennai, 2016, pp. 1-4. doi: 10.1109/ICCIC.2016.7919592
12. Maitrey, S. Jha, C. K. Gupta, R. and Singh, J. (2012). Article: Enhancement of CURE Clustering Technique in Data Mining. IJCA Proceedings on Development of Reliable Information Systems, Techniques and Related Issues (DRISTI 2012) DRISTI(1):7-11, April 2012.
13. Garima, Gulati, H. and Singh, P. K. (2015). Clustering techniques in data mining: A comparison, 2nd International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, 2015, pp. 410-415.
14. Tiruveedhula, S. Rani, C. M. S. Venkata, C. M. N. (2016). A Survey on Clustering Techniques for Big Data Mining, Indian Journal of Science and Technology, [S.l.], feb. 2016. ISSN 0974 -5645. doi:10.17485/ijst/2016/v9i3/75971

AUTHORS PROFILE:

Nikita Kumble is an M.Tech student in the computer Engineering department of S.G.S.I.T.S, Indore. She has received a Bachelor of Engineering degree Rajiv Gandhi Proudyogiki Vishwavidyalaya, Bhopal(2016). Her research interest includes data mining, Machine Learning and Soft Computing .

Vandan Tewari is an Associate Professor in the Department of Computer Science Engineering at Shri Govindram Seksaria Institute of Technology and Science, Indore. She earned her doctoral degree from Rajiv Gandhi Proudyogiki Vishwavidyalaya, Bhopal. Her research interest includes Data Mining, Databases, Data Science and Social Network Mining.