Scientific Journal of Impact Factor (SJIF): 3.134

ISSN(Online): 2348-4470
ISSN(Print) : 2348-6406

International Journal of Advance Engineering and Research Development

Volume 1, Issue 11, November -2014

AN ENHANCED ALGORITHM FOR IMPROVED CLUSTER GENERATION TO REMOVE OUTLIER'S RATIO FOR LARGE DATASETS IN DATA MINING

Mayuri G. Vadgasiya[1], Prof. Jay M. Jagani[2]

[1]*M.E. [Computer Engineering], Darshan Institute of Engineering & Technology, Rajkot, patelmayu18@gmail.com*
[21]*M.Tech. [Computer Engineering], Darshan Institute of Engineering & Technology, Rajkot, Jay.jagani@darshan.ac.in*

Abstract *— existing studies in data mining focus on Outlier detection on data with single clustering algorithm mostly. There are lots of methods available in data mining to detect the outlier by making the clusters of data and then detect the outlier from them. Outlier can be reduced if we improve the clustering. The values or objects that are similar to each other are organized in group it's called cluster and the values or objects that do not comply with the model or general behavior of the data these data objects called outliers. Outliers detect by clustering. We make algorithm that will be generate the percentage value of cluster and the outliers and its compulsory to total no of cluster percentage are greater than the total no of outlier percentage. If the cluster are not more than outliers then algorithm will be improved the total no of cluster and reduce the outliers. The output of the algorithm will be generating total original objects. If the no of input objects and no of output objects are not same then we assume that some error occur in the algorithm.*

Keywords*-Data mining; Clustering; Outliers; Clustering algorithm; Hierarchical Clustering Algorithm*

I. INTRODUCTION

Clustering can be used to generate class labels for a group of data. The objects are clustered or grouped based on the principal of maximizing the interclass similarity and minimizing the interclass similarity. That is, cluster of objects are formed so that objects within a cluster have high similarity in comparison to one another, but are rather dissimilar to object in other clusters. This technique designed as undirected knowledge discovery or unsupervised learning. There are lot of clustering techniques which are used to generate the clusters from data[1].

The objects and values that do not comply with the model or general behavior of the data these data objects called outliers. Many data mining methods discard outliers as noisy or exceptions. Outliers is also observed that the davits of other observations are behaves like arouse suspension and it was generated by different mechanism [1].

Outlier detection is an important research problem that aims to find objects that are considerably dissimilar, exceptional and inconsistent in the large amount of database [4]. And the Detection of such outliers is also important for many applications such as fraud detection and customer migration [2]. In this paper we enhanced the algorithm that will be derive the percentage of the clustered object and the outlier's object. And it also checks that the total no of percentage of the cluster object is greater than the total no of percentage of outlier object. If the total no of object of the cluster are less than outliers then improve cluster generation to remove outlier's ratio. And it also checks the algorithm that will give same no of input object in output. If the total of outliers and clustered dataset will be the original input dataset then one can say that there will be error either in clustering or in outlier detection. By improving cluster generation on large dataset that will reduce outliers and it will give improved performance and accuracy of the sum of outliers and clustered data nearby same as input data.

II. LITERATURE

Clustering algorithm can be divided in various categories. Following are the categories of clustering algorithm.

Types of Cluster Algorithms

1) Partition Clustering Algorithm
2) Hierarchical clustering
3) Density Based Algorithm
4) Grid based Method
5) Model Based Method

2.1. Partition Clustering Algorithm

Partition clustering algorithm splits the data point into K partition, every and each partition can represent a cluster. When the partition is done based on certain objective releated function.

Partitioning clustering [3] attempts to decompose a set of N objects into k clusters such That the partitions optimise a certain criterion function. Each cluster is represented by the centre of gravity (or centroid) of the cluster, e.g. k-means, or by the closest Instance the closest instance to the gravity centre (or medoid), e.g. k-medoids. Typically, k

seeds are randomly selected and then a relocation scheme iteratively reassigns points between clusters to optimise the clustering criterion. The Minimisation of the square-error criterion [5] sum of squared Euclidean distances of points from their closest cluster representative point, is the most commonly Used.

Formally, given a data set, D, of n objects, and k, the number of clusters to form, a Partition clustering algorithm splits the data point into k partition, every and each partition can represent a cluster**.** The clusters are optimize an objective partitioning criterion, such as a dissimilarity function based on distance so that the objects within a cluster are similar to one another and dissimilar to objects in other clusters.

2.1.1. K-Means: A Centroid-Based Technique

K-means is perhaps the most popular clustering method in metric spaces [3,5,15,16]. Initially k cluster centroids are selected at random. K-means then reassigns all the points to their nearest centroids and recomputes centroids of the newly assembled groups. The iterative relocation continues until the criterion function, e.g. square-error, converges. For its wide popularity k-means is sensitive to outliers and noise since a small number of data are substantially affects the centroids**.**

2.1.2. K-Medoids Algorithm

Unlike k-means, in the k-medoids or Partitioning around Medoids (PAM) [3,17] method a cluster is represented by its medoid that is the most centrally located object (pattern) in the cluster. Medoids are more resistant to outliers and noise compared to centroids. PAM begins by selecting randomly an object as medoid for each of the k clusters. Then in medoids each of the non selected objects are grouped with which it is the most similar. PAM then iteratively replaces one of the medoids by one of the non-medoids objects yielding the greatest improvement in the cost function. There are other partitioning algorithms such as K-modes and FCM.

2.2. Density Based Clustering Method

In density-based methods, outliers are detected from local density of observations. Purpose of this algorithm to be discovering areas of high density that are separated from each other by area of low-density based. A low density of the observation is an indication of a possible outlier [8]. Density-based methods have noise tolerance, and can discover nonconvex clusters. Similar to hierarchical and partitioning methods, density-based techniques encounter difficulties in high dimensional spaces because of the inherent sparsity of the feature space, which in turn, reduces any clustering tendency [3]. DBSCAN and DENCLUE are algorithms that use such a method to filter out outliers (Noise) and discover clusters of arbitrary shape.

2.3. Grid Based Method

The space of the Grid based method [9] is divided into grids. Fast processing time is the main advantage of this method because to compute the statistical values it goes through the dataset once for the grid**.** The accumulated grid-data make grid-based clustering techniques independent of the number of data objects that employ a uniform grid to collect regional statistical data, and then perform the clustering on the grid, instead of the database directly. The performance of a method depends on the size of grid, and that size is usually much less than the size of database. However, for distributions of highly irregular data single uniform grid is not be sufficient to fulfill the time requirement and obtain the required clustering quality. OPTIGRID is examples of this category.

2.4. Model Based Method

Model based clustering method [7] construct the model for every clusters and find a data which is fit to that model and this method is automatically give the number of clusters. This method is robust. The representative model-based clustering algorithm is EM. There are two major approaches that are based on the model-based method: statistical and neural network approaches. Model based method is often based on probability distribution of data. Individual distribution is called component distribution .in this method probability distribution is done by the mixture density model. EM method acquire statistic from traditional mixture model and depends on that statistic it perform clustering in model based clustering method. There are other Model based algorithms such as EM, SOMs.

III. HIERARCHICAL CLUSTERING

A Hierarchical clustering Method [5] is a one of the procedure that will be used for transforming a proximity matrix into a sequence of nested partitions. For performing a Hierarchical clustering the Hierarchical clustering algorithm is used as a specification of steps. It is often acceptable to characterize a Hierarchical Clustering method by writing down an algorithm. But the algorithm should be separated itself from the method. A Hierarchical Clustering is a sequence of partition in which each partition is nested into the next partition in the sequence. Hierarchical Cluster builds a cluster Hierarchy or, in other words, a tree or cluster, also known as dendrogram [6]. Every cluster node contains child clusters, sibling cluster partition the points covered by their common parent [6]. Such an approach allows exploring data on different level of granularity [6].

Advantage of Hierarchical Clustering [6]:

- Irrespective the level of granularity flexibility is embedded.
- In any forms of similarity or distance Easy to handling.

Disadvantage of Hierarchical Clustering [6]:

- The fact that most hierarchical algorithms do not revisit once constructed (intermediate) clusters with the direction of their improvement

3.1. Hierarchical Clustering Method
There are two types of hierarchical clustering method agglomerative (bottom- up) and divisive (top-down) [6].

3.1.1. Process of agglomerative:
The agglomerative method [12], also known as the bottom-up method, and that will be starts with the entire object forming a separate group. It merges the objects or groups that are close to one another, until all the groups are merged into one (the topmost level of the hierarchy), or a termination condition holds. When cluster with different size in the tree can be valuable for discovery process of agglomerative. An agglomerative clustering starts with one-point (singleton). Clusters and recursively merges two or more most appropriate clusters.
Step 1) allocated every object to a separate cluster.
Step 2) Classify all-pair-wise distances between all clusters.
Steps 3) build up a distance matrix.
Step 4) Select the pair which is distance is very shortest.

- Obliterate the pair from the matrix form and it merges them.
- Classify again all distance from the new cluster to all other cluster.
- When update previous matrix.
- This process can be repeat unit the distance matrix is reduced to a single element.

3.2.2. Process of Divisive:
The divisive method [12], also known as the top-down method, and that will be starts with all the objects in the same cluster. In each successive iteration, a cluster is split into smaller clusters, until each object is to be reformed in one cluster, or a termination condition holds**.** Always to be large cluster are successively divide. A divisive clustering starts with one cluster of all data points and recursively splits the most appropriate cluster.
Step 1) initially start with all object in one cluster C
Step 2) when divide successively each cluster C into smaller divide (C1, C2).

- When each variable y, find the cutting s, which optimizes

$$W(c) = q(c1) + q(c2)$$

- Choose the variable y and the cutting s which optimizes w(c).

Step 3) Divide the cluster c→ (c1, c2), which is maximizes,

- $A(c) = | q(c)-q(c1)-q(c2) |$

Step 4) END

3.2. Hierarchical Clustering Algorithm
There are many types of Hierarchical Clustering algorithms such as BIRCH, CURE, and CHAMELEON.

3.2.1. BIRCH (Balanced iterative Reducing and clustering using Hierarchies):
Balanced Iterative Reducing and Clustering Using Hierarchies (BIRCH) [3,18**]** introduces a novel hierarchical data structure, CF-tree, for compressing the data into many small sub-clusters and then performing clustering with these summaries rather than the raw data. BIRCH is local in that each clustering opinion is made without scanning all information points. While scanning the database the BIRCH algorithm consider those data points which are near to each other. Sub-clusters are represented by compact summaries, Called cluster-features (CF) that are stored in the leafs. The non-leaf nodes store the sums of the CFs of their children. A CF-tree is built dynamically and incrementally, requiring a single scan of the dataset. An object is inserted in the closest leaf entry. Two input parameters control the maximum number of children per non-leaf node and the maximum diameter of sub-clusters stored in the leafs. By varying these parameters, BIRCH can create a structure that fits in main memory. Once the CF-tree is built, any partitioning or hierarchical algorithms can use it to perform clustering in main memory.
Advantage of BIRCH:
1) BIRCH requires a single scan of the database
2) BIRCH algorithm is proposed in the database are easy handle noise effectively.
3) That method does not required the whole dataset in advance because It is an Incremental Method.

3.2.2. CURE (clustering using representative):
Clustering Using Representatives (CURE) is an agglomerative method. CURE clustering algorithm is a new hierarchical algorithm basically it can be adopts in a middle level between centroid based and all point. CURE is an efficient clustering algorithm for large database is more robust in outliers to be identifying non-spherical shapes of cluster and wide variances cluster size. Cluster with close pair of representative point of cluster it can be merged each and every step of cure. At each iteration, the Pair of clusters with the closest representatives is merged [3]. Random sampling and

partitioning technique can be used for reduce the input dataset. CURE uses a combination of random sampling and partitioning to improve scalability [3].

Advantage of CURE:

1) CURE is a clustering algorithm for large database and is more robust in outliers to be identify non-spherical shapes of cluster and wide variance cluster size

3.2.3. CHAMELEON (Multiphase Hierarchical Clustering Using Dynamic Modeling)

CHAMELEON [3,19] improves the clustering quality by using more elaborate merging criteria compared to CURE [3,19]. Initially, a graph containing links between each point and its k-nearest neighbors [3,5] is created. Then a graph-partitioning algorithm recursively splits the graph into many small unconnected sub-graphs. During the second phase, each sub-graph is treated as an initial sub-cluster and an agglomerative hierarchical algorithm repeatedly combines the two most similar clusters. Two clusters are eligible for merging only if the resultant cluster has similar inter-connectivity and closeness to the two individual clusters before merging. Due to its dynamic merging model CHAMELEON is more effective than CURE in discovering arbitrary-shaped clusters of varying density**.** After all better performance comes at the budget of computational cost that is equalizing in the database size.

Advantage of CHAMELEON:

1) CHEMELEON has greater power for discovering arbitrarily shaped clusters of high Quality than well- known algorithms like BIRCH and DBSCAN.

IV. COMPARISION TABLE

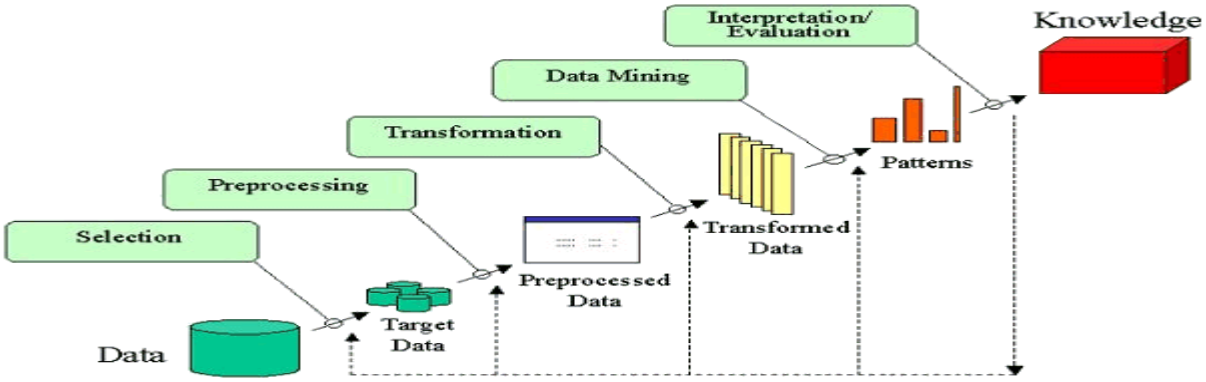
The Comparison Table [9] of the Clustering Algorithm is as follows:

Categories	method name	Size of Data-set	Handling High Dimensionality	Handling Noisy Data	Type of Dataset	Input Parameter
Partitional algorithms	K-Means	Large	No	No	Numerical	1
	K-medoids	Small	Yes	Yes	Categorical	1
Hierarchical algorithm	BIRCH	Large	No	Yes	Numerical	2
	CURE	Large	Yes	Less Sensitive	Numerical	2
	CHEMELEONE	Large	Yes	No	All type data	3
Density Based Algorithm	DBSCAN	Large	No	No	Numerical	2
	DENCLUE	Large	Yes	Yes	Numerical	2
Grid Based Algorithm	OptiGrid	Large	Yes	Yes	Special Data	3

V. ROLES OF OUTLIERS

The affection of ascertain ability is totally dependent on data affection. On the other hand actual data uses unwanted noise, ambiguity, errors, overabundance or unsuitable data. The more complicated thing is the actuality to analyze; the big problem is getting down affection data. Ability ascertains from databases (KDD) offers a all-inclusive framework to anticipate data in the appropriate form to perform actual analyses. Additionally, the affection of adjustment taken affixed to KDD outcomes, satisfy not only on the affection of the outcomes themselves, but it is reachable to communicate those outcomes in an understandable form [11].

Knowledge Discovery of Data (KDD) [11] perform in 1989 referring to high level applications in which include accurate methods of Data Mining (DM, see figure) to select Useful and understandable knowledge from data. KDD processes and the application of DM Techniques are specifically unfaithful for environmental data, since activities permitting extraction of maximum useful information from data bases are very important although they use to be before for an environmental software system development. Also the KDD approach facilitates the combination of different knowledge sources and fields of ability and the involvement of end-user (domain expert) criteria and stakeholder's points of view in algorithm design and result interpretation. Finally, it facilitates the sharing and active re-use of data and extracted technical knowledge and experiences among domain experts.



Outliers that will be detected from the preprocessing state of the KDD Process. Outliers are objects with very extreme values in one or more variables [11] Graphical techniques were once realistic approach for understanding them, but incremental value in database sizes and dimensions point out to a variety of automated techniques. The use of standard deviations is accessible when and only when reviewing a single variable that has an equal assessment, but outlier may also take the form of abnormal connection of two or more variables. The data point should be analyzed as a whole to understand the nature of the outlier and variable method is required.

VI. PROPOSED WORK

There are lots of methods are available to improved a clustering ratio and detect a outliers ratio but all the methods have a limitations so any one method cannot overcome all the parameters like Geometry shape, Outlier Handling, Running time, Noise handling, Complexity, Large Dataset Handling. all the algorithm have advantage over any other algorithm so if we combined any two algorithm then the result will be more better than the independent algorithm.

BIRCH and CURE are the Hierarchical algorithms both algorithms are used for a large dataset but CURE algorithm handle large dataset more effectively. BIRCH and CURE both algorithm handle a noise but BIRCH can handle a noise effectively where as CURE is less sensitive to handling a noise. BIRCH algorithm can identifies only spherical or convex shape of cluster where as CURE algorithm can identifies a arbitrary shape. so if we combined a BIRCH and CURE algorithm then we got a hybrid algorithm and that will be overcome all the advantage of the BIRCH and CURE. The result of that hybrid algorithm is much better than the independent BIRCH and CURE algorithm.

VI. CONCLUSION

For a improve cluster or outlier detection many types of algorithms are available but all the algorithms have some limitation and only one algorithm cannot give perfect output so for improve an output we have need to combined any two algorithm they have advantage over each other. BIRCH can handle a noise more effectively compared to CURE. BIRCH Algorithm can identify only convex or spherical shape whereas CURE can identify arbitrary shape so BIRCH and CURE both have some advantage over each other if we combined BIRCH and CURE then we got improved algorithm and that algorithm will be overcome all the advantage of BIRCH and CURE.

REFERENCES

[1] Nancy Lekhi, Manish Mahajan "Improving Cluster Formulation to Reduce Outliers in Data Mining" International Journal of Innovative Research in Computer and Communication Engineering Vol. 2 Issues 6, pp. 13–15 , June- 2014.

[2] Zengyou He, Xiaofei Xu, Shengchun Deng "An Optimization Model for Outlier Detection in Categorical Data" Department of Computer Science and Engineering Harbin Institute of Technology, pp. 1-3.

[3] Ioannis Sarafis "Data Mining Clustering of High Dimensional Databases with Evolutionary Algorithms" pp. 24-29, August 2005.

[4] Navneet Kaur, Prof. Kamaljit Kaur " Comparison between two Approach Based on Threshold and Entropy Based Approach " International Journal of Advanced Research in Computer Science and Software Engineering Vol. 3 Issue 8, pp. 81-84, August- 2013.

[5] Anil K. Jain, Richard C. Dudes "Algorithms for Clustering Data" Prentice Hall, A division of simon & Schuster Englewood Cliffs, New Jersey 07632, pp. 58-92.

[6] Pavel Berkhin "Survey of Clustering Data Mining Techniques" Accrue Software, Inc., pp. 6-12.

[7] Karuna Katariya, Rajanikanth Aluvalu "Agglomerative Clustering in Web Usage Mining: A Survey" International Journal of Computer Applications (0975 – 8887)Vol. 89 Issues 8, pp. 25-27, March- 2014.

[8] Professor Pasi FrÄanti, Professor Martti Juhola, Professor Olli Nevalainen, Professor Pekka KilpelÄainen "Improving Pattern Recognition Methods for Speaker Recognition" Joensuu, pp. 17-23, 2008.

[9] A. Fahad, N. Alshatri, Z. Tari, Member, IEEE, A. Alamri, I. Khalil A. Zomaya, Fellow, IEEE, S. Foufou, and Bouras "A Survey of Clustering Algorithms for Big Data: Taxobomy & Empirical Analysis", pp.1-3, 2014.

[10] Joaquín Izquierdo, Geoff Holmes, Ioannis Athanasiadis, Joaquim Comas, Miquel Sanchez-Marre "On the role of pre and post-processing in environmental in data- Mining" International Environmental Modelling and Software Society (iENSs), 2008.

[11] Barnett, V. and Lewis, T. 1978. "Outliers in Statistical Data" Wiley.

[12] KAUFMAN, L. and ROUSSEEUW, P. 1990. "Finding Groups in Data: An Introduction to Cluster Analysis" John Wiley and Sons, New York, NY.

[13] J. J. Fortier and H. Solomon. "Clustering procedures. In Multivariate Analysis" Pages 439–506, 1966. New York.

[14] R. E. Jensen. "A dynamic programming algorithm for cluster analysis" Operations Research, 17, pp.1034–1057, 1969.

[15] J. Hartigan and M. Wong. "A k-means clustering algorithm" Applied Statistics, pp:100–108, 1979.

[16] J. MacQueen. "Some methods for classification and analysis of multivariate Observation In Proc" 5th Berkeley Symp. Math. Statist. Prob., pp: 281–297, 1967.

[17] L. Kaufman and P. J. Rousseeuw. "Finding Groups in Data: An Introduction to Cluster Analysis" John Wiley & Sons, 1990.

[18] T. Zhang, R. Ramakrishnan, and M. Livny. "BIRCH: an efficient data clustering Method for very large databases" In Procc of the ACM SIGMOD Int. Conf. on Management of Data, volume 25, pp: 103–114, 1996.

[19] G. Karypis, E-H. Han, and V. Kumar. "Chameleon: Hierarchical clustering using dynamic modeling" IEEE Computer, 32(8) , pp: 68–75, 1999.

[20] R. Ng and J. Han. Efficient and effective clustering method for spatial data mining. In Proc. of 1994 Int. Conf. on Very Large Data Bases (VLDB'94), pp:144–155, 1994.

[21] Hans-Peter Kriegel, Martin Pfeifle," Density-Based Clustering of Uncertain Data"

[22] M. Ester, H-P Kriegel, J.Sander, and X. Xu. "A density-based algorithm for discovering clusters in large spatial databases with noise" In Second International Conference on Knowledge Discovery and Data Mining, pp:226–231, 1996.

[23] A. Hinneburg and D. A. Keim. "An efficient approach to clustering in large multimedia databases with noise" In Proc. 1998 International Conference on Knowledge Discovery and Data Mining (KDD'98), pp: 58–65, 1998.

[24] D. W. Scott. Multivariate Density Estimation. Wiley, New York, 1992.

