IJCSE International Journal of Computer Sciences and Engineering Open Access

Review Paper Volume-4, Issue-3 E-ISSN: 2347-2693

A Review of Clustering Methods forming Non-Convex clusters with, Missing and Noisy Data

Sushant Bhargav[1*] and. Mahesh Pawar[2]

[1]*School of Information Technology, RGPV, India*
[2]*Department of IT, UIT, RGPV, India*

www.ijcseonline.org

Received: Feb/26/2016 Revised: Mar/06/2016 Accepted: Mar/17/2016 Published: Mar/31/ 2016

Abstract— Clustering problem is among the foremost quests in Machine Learning Paradigm. The Big Data sets, being versatile, multisourced & multivariate, could have noise, missing values, & may form clusters with arbitrary shape. Because of unpredictable nature of Big Data Sets, the clustering method should be able to handle missing values, noise, & should be able to make arbitrary shaped clusters. The partition based methods for clustering does not form non-convex clusters, The Hierarchical Clustering Methods & Algorithms are able to make arbitrary shaped clusters but they are not suitable for large data set due to time & computational complexity. Density & Grid Paradigm do not solve the issue related to missing values. Combining different Clustering Methods could eradicate the mutual issues they have pertaining to dataset's geometrical and spatial properties, like missing data, non-convex shapes, noise etc.

Keywords— *Clustering, convex, non-convex, missing values, Big Data, noisy data, data mining, density based*

I. INTRODUCTION

Clustering can be defined as grouping of data on the basis of similar attributes or properties of subsets of data from a dataset, which can be used for distinguishing data groups as clusters. In present era applications generate data that only accumulate more and more data. This data are contributed by Internet services, scientific experiments, Sensors, Flights, Industries, News, and Media etc. For example Social Media like Facebook generates 500 TB of data each day, YouTube gets 72 hours of video upload each day by its users, and the list goes on.

Figure 1 shows a graphical relation between the growths of data generation rate per month from 2014 to 2019. Annual world Information Processing traffic might cross one thousand Exabyte mark by December of 2016, and may reach two Zettabyte per annum by 2019. By 2016, world IP traffic can reach a yield of 1.1 Zettabyte annually, or about one billion GB per month, international IP traffic may yield two Zettabyte data each year, or 168 Exabyte per month [1].

Big volume of data requires management and reuse as well, so that data or some analytical aspects of data can be reused to prevent reinvention of wheel. This enormous quantity of data can be reused by corporations and government for analysis & save significant time by preventing recollection of the same data, however a large volume data or big data have its own issues. Large volume of data demands large space and heavy computing, like analytical operations, fetch and retrieve operations, Processes, etc. So instead performing operation on entire Big Data, we perform an operation on an image of data which projects distribution of values among real data, such images or views of data are called Clusters [2]

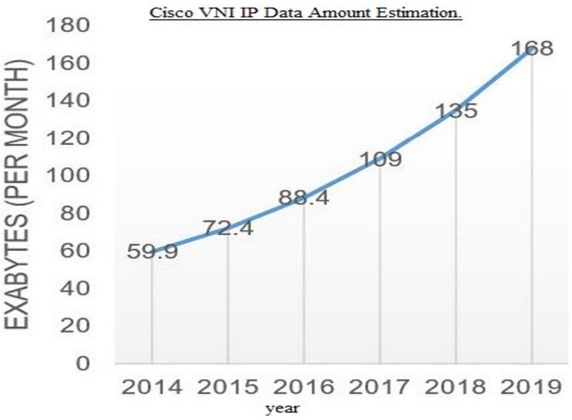

Figure 1 Data growth per month by 2019 [1]

As the quantity of data will grow so will the need of algorithms that can analyze data without any presumptions, like K-Means which insists on Convexity of data, and Algorithms should also be robust, scalable, capable of handling Big Data and not susceptible to inconsistencies like missing data.

Corresponding Author: *Sushant Bhargav, BhargavSushant@gmail.com, SOIT, RGPV, India*

Since, the data that is being accumulated is accounted from various sources, even sources like Large Software Development IDEs, Databases and Mathematical Engines & Weblogs [17], the data is expected to be of different nature (called variety) & sometimes the sources are not accounted for or likely to be error prone, like sensors, such data is expected to have uncertainties like missing values, noise etc., this property that references uncertainty of data is called Veracity.

There are many Clustering algorithms in data mining that are very efficient at grouping data, major issues with clustering problem is that algorithmic & time complexity increases exponentially with respect to dataset size [3]. In real world scenario the data may have multivariate attributes, missing values, noise, non - convex etc.

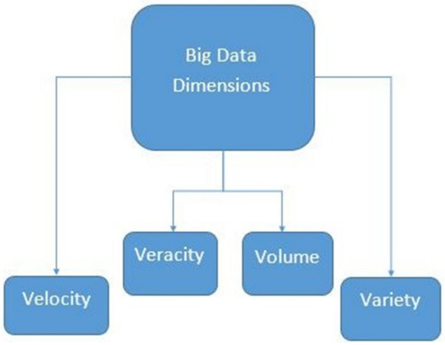

Figure 2 Big Data Dimensions

A big data clustering algorithm should be able to handle a variety of data efficiently. In this paper, we survey various algorithms for clustering and their working on different types of datasets. Clustering algorithms can be broadly categorized as Partitioned Based, Hierarchical, Density Based, Grid Based and Model Based. (Fahad et al. 2014).

This paper has been organized into following parts:

Section II. Literature review, III. Analysis, IV. Conclusion, & V. References.

Part II, contains literature review, which is a brief review of the literature available on clustering non convex data and popular clustering methodologies and practices. Part III is analysis of various clustering methods and their usefulness pertaining to the given problem statement, i.e. Clustering Non Convex data with missing information and noise. In Part IV we summarized conclusion of our study and in Part Vth the appropriate references are given to cite the sources we used in this study.

II. Literature Review

In [4] K-Means are considered as one of the simplest algorithms in machine learning to solve the clustering problem. In this algorithm number of clusters are predefined as 'K', which refers to the number of centroids. The algorithm

Iteratively calculates the closeness between points to each centroid and as a result compact groups of items are formed as clusters. K-Means uses squared error function as the objective function, and it minimizes the squared error distance between items and chosen centroid.

In Partition based clustering, a dataset is considered to have a fixed number of clusters, the objective of a partition-based algorithm is to divide the set of objects into a set of predefined k- disjoint clusters where k is the number of clusters.

Partition based algorithms are iterative & use a distance formula to measure similarity between items. The main advantage of such algorithms is that iteration to create the clusters, the drawback of these algorithms are that they need to have a value of k beforehand, they are susceptible to outliers and noise and also these algorithms cannot make non-convex clusters [5].

In [7] & in [8], it is mentioned that the objective function of kernel k means and that of spectral clustering is same, so it implies that both spectral clustering and kernel k means are able to make non convex clusters.

It is mentioned in [7]that Ensembling enables formation of non-convex like shape, that by taking multiple looks at the same data, by generating multiple partitions (also called clustering ensemble) of the same data.

At the end the results are combined to form clusters that are very close or not well partitioned, like non convex shapes.

In Hierarchical clustering, Data is organized according to hierarchy of proximity which are generated by the intermediate nodes. Entire dataset can be represented by Dendrogram in which the leaf nodes represent the data itself. Like a Tree, A Data-Class can have a subclass and this division can continue to leaf. There are two types of such clustering divisive (top-down) & agglomerative (bottom-up).

Figure 3 is a result of K-Means algorithm on Jain's toy dataset. [6] which is drawn using R Studio [16]

In [9], the authors presented a new clustering algorithm using hierarchical clustering with convex relaxation, this is done to generate a set or a class of objective functions that have more casual or natural geometric interpretation. This algo also can perform learning by establishing tree structure of data and can generate non convex clusters like spectral clustering.

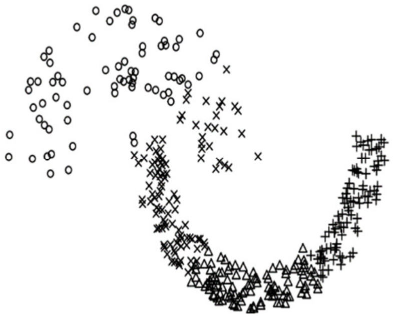

Figure 3 non convex cluster result by K-Means

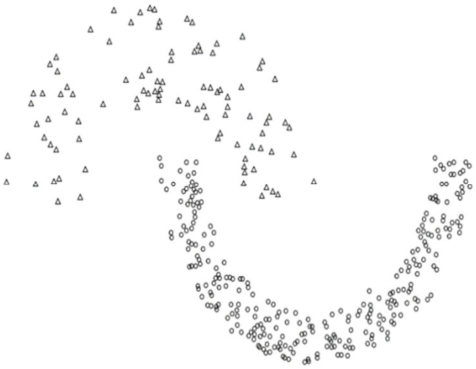

Figure 4 Expected cluster on Toy Dataset

In [10] an algorithm is proposed to calculate clusters on the basis of density, connectivity and boundary, clusters are made with a clear separation of density in region, the Density based algorithms can find clusters that have arbitrary shapes, like Jain's toy dataset, or double moon dataset, etc. The results of DBSCAN experiments conclude that
(1) It is more effective in measuring and determining cluster parameters and cluster shape than CLARANS,
(2) Efficiency of DBSCAN is greater than that of CLARANS by a factor of 100 times, and also it can make non convex clusters. However DBSCAN doesn't deal with missing data problem, but it can handle outliers.

Figure 5 is a result of DBSCAN algorithm on A.K.Jain's Toy data set [6], it is a result of a standard DBSCAN on Jain's Toy Dataset is shown, it is clear that DBSCAN is able to perform non convex clustering.

In [11] a novel framework is proposed, which combines Density approach, grid-based clustering algorithm & sliding window model. The proposed algorithm is named DEN-Gris, combining the Density and Grid approach. It combines the best feature of DBSCAN and puts a capping of time and memory on the algorithm to increase performance, which is very vital when using on large data sets. The DENGRIS algorithm has an Online Component and an Offline Component, Online component maps each data record to a density grid per sliding window, the space of the data objects is divided into grids. The offline Component, on the other hand performs clustering, by removing sparse grids, then merging the neighboring denser grids.

The DENGRIS is more suitable for large data set, as it has different learning components, it is immune to outliers, and can form non convex clusters like DBSCAN, however it does not deal with missing values as well. It has faster processing because it calculates values for the data only once, then it stores the value onto a grid, thereby making a *regional statically database*, hence the clustering is performed on the grid rather than the dataset, which saves time for modeling the data once again. This algorithm may not work efficiently if there are too many of outliers or missing data.

[12] In this paper spectral clustering is explained by creating a similarity graph based on Eigenvalue matrix, and clustering is done on the basis of compactness, graph Laplacian enables us to make clusters that are not necessarily compact.

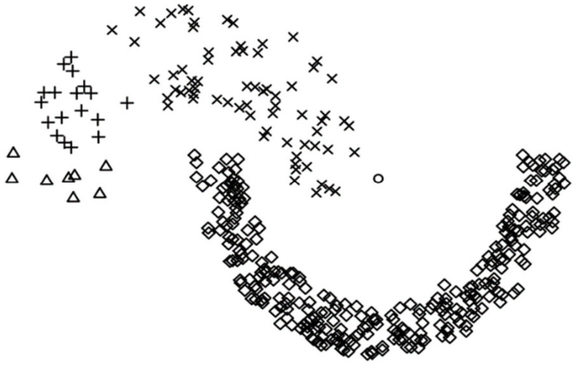

Figure 3 DBSCAN result on Toy Dataset

In [13] Graph theoretic clustering in conjunction with EM and Rough Set Theory is used, by using the EM instead of real values clustering is performed on Gaussians,.
A Gaussian mixture model is most helpful in making clusters in case of noise and outliers, Rough Set Theory helps EM to not settle on local minima, and Graph Theoretic Clustering uses a Minimum Spanning Tree for Clustering. Graph Theoretic Clustering tackles problem of non-convex clusters.

Model based methods optimize fit between the given data and a predefined mathematical model. EM assumes that Data is yielded by a mixture of the underlying probability distribution. Model based approaches are broadly divided into two categories, Statistical and Neural. Statistical methods use Probably Distribution Model to analyze the dataset, best of Model based approaches are Mclust and Cobweb. Neural Network clustering performs very well ANN based clustering algorithms take data points as input and adjusting and re-adjusting their weights to extract patterns from datasets, this is called learning of ANN. The Exemplar is one of the ANN methods used for clustering.

In [14] it is concluded that partitioning based clustering is fit for spherical shaped or convex clusters & for small to medium sized data sets. K-Medoid contrary to K-Means is immune from outliers and noise, but still needs the value of K in advance and it cannot form non-convex clusters.

Table 1 Comparison of Studied Clustering algorithms

	Noise Immunity	Missing Values Handling	Non-Convex clustering
K-Means	No	No	No
K-Medioid	No	No	No
Kernel K-Means	No	No	Yes
Spectral Clustering	No	No	Yes
Hierarchical Clustering	No	No	Yes
Graph Theoretic Clustering	No	No	Yes
DenGris	No	No	Yes
DBSCAN	No	No	Yes
EM	Yes	Yes (By Imputation)	No
EMG	Yes	No	Yes

In this part we studied about various clustering algorithms, Partitioning based clustering algorithms like K-Means are fast and popular but unable to deal with arbitrary shaped clusters. Model Based Algorithms require complex mathematical model to fit on the existing dataset in order to drive out any conclusive patterns. Grid Based algorithms form a structural relation between data and thus very time consuming at the beginning but provide faster result later, Density based algorithms are best for non-convex data but slow on large datasets, and Hierarchical Clustering algorithms can make non convex shapes but then they can be very slow on large datasets.

Table 1 compares a concise result of this study, we compare the algorithms based on their ability to Immunity to noise, Handling Missing Values, and Non-Convex clustering.

III. Analysis

- Functioning of clustering algorithms depends on the mathematical & logical procedure which they use to solve a problem, and also on the input dataset. Partition based clustering algorithms address dataset with convex grouping or patterns. Density based clustering algorithms basically use eps distance and minimum points to designate a group of points as cluster. Hierarchal based clustering is used where there is need of classification among the data points on the basis of hierarchy. Model based clustering algorithms use mathematical model as a fit on dataset in order to find meaningful patterns. Grid based clustering algorithms form grid on dataset in order to find patterns or clusters.

- There are shortcoming to almost every clustering algorithm because they are devised keeping an input pattern in mind, and the procedure is intended to form clusters that can resemble original pattern as closely as possible.

- Spectral Clustering, DBSCAN, EMG, GTC, DenGris are able to perform non convex clustering but do not deal with missing values, so by combining above algorithms with a method that can estimate missing values, an algorithm can be invented that is versatile enough to handle Missing values, Noise and Non Convex Clusters.

- Hierarchical clustering algorithms are not suitable for Large Dataset due to time complexity.

IV. Conclusion

- In this paper we studied various clustering methods that address issues related to clustering such as Non Convex data, Noise, Missing values, Speed etc.
- Partition based clustering algorithms are fast and easy to implement, but face problems with inconsistency in data and depend on parameter initialization. However, there are algorithms like Kernel K Means and methods

that can combine both K-Means and other algorithms which are not restricted to convex shaped clustering, can make arbitrary shaped clusters, like combining K-Means and then agglomerating the resultant clusters using Hierarchical Clustering algorithms can yield non-convex shapes. Although such methods are costly in terms of processing as Hierarchical clustering are not suitable for clustering Big Data Sets due to time complexity, and it does not eradicate the need of initialization of cluster parameters, & effects of noise and missing data.

- Algorithms which can address different clustering related issues can be combined to address individual or a set of issues in order to eradicate each other's shortcomings, e.g. Combining K-Means and Hierarchical**.** Clustering can help to yield non convex dataset.
- Since K-Means can form clusters very effectively and faster than most algorithms, we can use this algorithm in convolution with Hierarchal clustering, by using Hierarchal clustering we can combine the patterns that are non-convex and then form clusters by using K-Means algorithm, like done in [15].
- Density based approach & Spectral clustering are also effective in making arbitrary shaped clusters, and able to handle noise and outliers however, it does not deal with missing data as well. Grid based approach, however if combined with density clustering, the resulting algorithm can yield non convex clusters and will be able to tackle noise, yet it cannot handle missing data.
- Finally we conclude that, though not every algorithm is designed to address all issues pertaining to clustering, specialized techniques have their own exclusive benefits, like K-Means forms Convex clusters efficiently and quickly, DBSCAN can form Non-Convex clusters but may perform slowly on large datasets, and may not yield good results with missing values.
- Therefore the convolution of multiple clustering algorithms is a popular method for solving clustering problem, there has not been much work done in forming unified clustering algorithms that can address clustering while a large dataset has missing values, noise, non-convex clustering & large datasets.

V. References

[1] Cisco, V. N. I. "The Zettabyte Era: Trends and Analysis." Updated :(Jun 23, 2015), http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/VNI_Hyperconnectivity_WP.pdf ; Document ID :1458684187584791 Accessed :Jan 2016

[2] Najlaa, Zahir, Abdullah, Ibrahim, Albert, Sebti, Bouras Fahad, "A Survey of Clustering Algorithms for Big Data: Taxonomy and Empirical Analysis," IEEE Transactions on Emerging Topics in Computing, vol. 2, no. 3, 2014.

[3] Leiserson, Rivest, Stein Cormen, Introduction to Algorithms, 3rd ed. ISBN 978-0262033848: Page 43-97, MIT Press & TMH, 2009.

[4] J.B.Macqueen, "Some Methods for classification and Analysis of Multivariate Observations," in 5th Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, Berkeley, 1967, pp. 281-297.

[5] Boomija, "Comparison of Partition Based Clustering Algorithms," Journal of Computer Applications, vol. 1, no. 4, p. 18, Oct-Dec 2008.

[6] A.K Jain and H.C. Martin, "Law, Data clustering: a user's dilemma," in In Proceedings of the First international conference on Pattern Recognition and Machine Intelligence, 2005.

[7] A.K.Jain, "Data clustering: 50 years beyond K-means," Pattern Recognition Letters, vol. 31, no. 8, pp. 651-666, June 2010.

[8] Vipin Kumar, Pang-Ning Tan, and Michael Steinbach, Introduction to data mining.: Addison-Wesley, 2005. ISBN : 9780321321367

[9] Joulin, Bach Hocking, "Clusterpath An Algorithm for Clustering using Convex Fusion Penalties," in 28th International Conference on Machine Learning , Bellevue, WA, USA, 2011.

[10] Martin, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu Ester, "A density-based algorithm for discovering clusters in large spatial databases with noise," in In Kdd, vol. 96, no. 34, 1996, pp. 226-231.

[11] Amineh, W. Ying Amini, "DENGRIS-Stream: A density-grid based clustering algorithm for evolving data streams over sliding window," in International Conference on Data Mining and Computer Engineering, 2012, pp. 206-210.

[12] Ulrike Von Luxburg, "A tutorial on spectral clustering," Statistics and computing, vol. 17, no. 4, pp. 395-416, 2007.

[13] Pabitra Mitra, Sankar K. Pal, and Aleemuddin Siddiqi, "Non-convex clustering using expectation maximization algorithm with rough set initialization," Pattern Recognition Letters, vol. 24, no. 6, pp. 863-873, 2003.

[14] Saline S Singh & N C Chauhan, "K-means vs K-Medoid: A Comparative Study," in National Conference on Recent Trends in Engineering & Technology, (NCRTET) BVM College, Gujarat, India, 2011.

[15] pafnuty.blog, By Aman Ahuja, Updated: (2013,

Aug) https://pafnuty.wordpress.com/2013/08/14/non-convex-sets-with-k-means-and-hierarchical-clustering/ Accessed :Jan 2016

[16] R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

[17] Chourasia, Richa, and Preeti Choudhary. "An approach for web log preprocessing and evidence preservation for web mining." (2014): 210-215.

AUTHORS PROFILE

Sushant Bhargav received a B.E. degree from RGPV University, India. He has more than two years of experience in programming and cloud technologies. He is Currently pursuing Masters in Technology in Information Technology from SOIT, RGPV, India. His research interests are in the area of Machine Learning, Big Data, Statical Analysis and Modeling & Data Mininng.

Dr. Mahesh K Pawar is Sr. faculty in Department of IT, UIT , RGPV, India . With 15 years of Academic Experience & three years of IT Industry Experience as a Software Engineer. His research interests are Software Engineering, Big Data, DBMS and Hadoop.