IJCSE International Journal of Computer Sciences and Engineering Open Access

Review Paper Volume-5, Issue-10 E-ISSN: 2347-2693

Clustering Algorithms – A Literature Review

B. Ramesh[1*], K. Nandhini[2]

[1*] Dept. of Computer Science, Chikkanna Govt. Arts college (Bharathiyar University), Tirupur, India
[2] Dept of Computer Science, Chikkanna Govt. Arts college, Bharathiyar University, Tirupur, India

*Corresponding Author: rameshbala50@gmail.com Tel.: +91-77080-05944

Available online at: www.ijcseonline.org

Received: 26/Sep/2017, Revised: 09/Oct/2017, Accepted: 20/Oct/2017, Published: 30/Oct/2017

***Abstract* -** Algorithms in data science are all the rage today with data scientists. With the right algorithm businesses can get measurable value from the huge volume of data collected. There are several types of algorithms- Regression, Clustering, Decision Tree etc. For this paper we will focus on clustering algorithms which are widely used in sorting and classifying big data. The way data is classified is critical to analysts studying the data to provide insights to business decisions. Every large data set can use clustering algorithms to process a variety of data to produce great results. Algorithms are used in image and data processing, calculations, and automated reasoning. The aim of this paper is to touch upon big data analytics, other authors views on this topic in our literature review section, define cluster analysis, present different clustering methodologies with its advantages and disadvantages, a comparison of different clustering algorithms, and wrap up with findings and discussion from our review papers.

***Keywords*—** Clustering, data mining, big data analytics

I. INTRODUCTION

Data Mining is the process of sorting through large data sets to gather useful information. During the mining process, computational software identifies patterns and establishes relationships to solve problems. Companies use this data to predict future trends. With the advent of 'Big Data' more extensive data mining techniques are needed because the size of the information is much bigger, the variety of data is wide-ranging and the speed at which it is being delivered is super-fast. There are several data mining techniques such as association, classification, clustering, organization, prediction, outlier analysis and such. Our focus is on clustering. Clustering loosely termed is a collection of objects 'similar' to one another and 'dissimilar' to objects of other clusters.

Clustering has its roots in mathematics, statistics and numerical analysis. Clustering is a method of unsupervised learning and used commonly for statistical data analysis in many fields like machine learning, data mining, pattern recognition, image analysis and bioinformatics [14]. One of the most dominant elements of cluster analysis is the choice of an appropriate similarity measure. The similarity measure selection is a data-dependent problem [3]. Analytics is the tool that provides companies detailed information on what happened, predict what will happen, and give understanding into what should be the action plan to be followed. Once the big data is split into clusters data is easier to analyse. The major terminology used in analytics is descriptive analytics, predictive analytics, and prescriptive analytics [17]. Descriptive analytics is analysis of historical data (which might be few seconds or few hours old). For example, in an oil well descriptive analysis means gathering and synthesizing several instruments to give conditions of the well in plain English. Predictive analytics uses more complex analytics to monitor production or drilling. By predicting machine failures in an oil field companies can save downtime costs and revenue losses. Prescriptive analysis includes everything from the predictive model but also includes a tailored breakdown. In a prescriptive model, it tells us how to adjust for events in the future. For example, using prescriptive analysis pairing real-time down-hole drilling data with production data of nearby wells can help adapt an oil producer's drilling strategy.

Rest of the paper is organized as follows. Section I contains the introduction of data mining, clustering, and analytics, Section II contains the literature review section of authors on different clustering techniques, Section III explains different clustering methodologies, Section IV describes results and discussions of papers reviewed, and Section V concludes research work with future directions.

II. LITERARTURE REVIEW

Triguero, I., Maillo, J., Luengo, J., García, S. and Herrera, F. present various improvements to the well-known data mining technique k-nearest neighbor's algorithm to come up with smart data. k-NN algorithm's weaknesses - noisy data and incomplete data have been addressed using Noise filtering and correction and missing values imputation models. Also through parallelism and data reduction the k-NN algorithm has become a core model to detect and correct imperfect data, eliminate noisy and redundant data, as well as correct missing values. They present several case studies that showcase k-NN algorithm as a unique model to obtain smart data from large amounts of potentially imperfect data.

Laloux, J.F., Le-Khac, N.A. and Kechadi, M.T. state that current distributed clustering approaches are predominantly generating global models by aggregating local result, hence losing important knowledge. They present a new distributed data mining approach where local models are not directly merged to build the global ones. Centralize clustering is carried out at each site (node) to build local models. These models are sent to the servers where clusters will be regenerated based on local models features. Considering how many corporations have geographically isolated data centers the authors objective is to reduce data collection expense by minimizing data communication and computational time, while getting accurate global results.

Halkidi, M. and Koutsopoulos, I. develop a novel approach for online distributed clustering of streaming data using belief propagation techniques. They use a two-level clustering approach to address the problem of clustering distributed streaming data. At the node level, a batch of data arrives at each time slot, and the goal is to maintain a set of salient data (local exemplars) at each time slot, which best represents the data received up to that slot. At each epoch, the local exemplars from distributed nodes are sent to the central location, which in turn performs a second-level clustering on them to derive a data synopsis global for the whole system. The local exemplars that emerge from the second level clustering procedure are fed back to the nodes with appropriately modified weights which reflect their importance in global clustering.

Liao, W.K., Liu, Y. and Choudhary, A. propose a grid-based clustering algorithm that employs the Adaptive Mesh Refinement (AMR) technique to address highly irregular data distributions. Instead of using a single resolution mesh grid, the AMR clustering algorithm creates different resolution grids based on the regional density and these grids comprise a hierarchy tree that represents the problem domain as nested structured grids of increasing resolution. Next, the algorithm considers each leaf as the center of an individual cluster and recursively assigns the membership for the data objects located in the parent nodes until the root node is reached. The team's experiments also showed the efficiency and effectiveness of the proposed algorithm compared to the grid-based methods using single uniform meshes. Since it is a grid-based method, it also shares the common characteristics of all grid-based methods, such as fast processing time, insensitivity to the order of input data, and the ability to separate real data from noise.

Fernandez, J.R. and El-Sheikh, E.M. state that with today's generation of high speed data streams traditional clustering and/or pattern recognition algorithms are inefficient for clustering data. They define data stream as a dynamic dataset that is characterized by a sequence of data records that evolves over time, has extremely fast arrival rates and is unbounded. In their paper, they present a clustering framework (CluSandra) and algorithm that, combined, address the time constraint and space challenge, and allows end-users to explore and gain knowledge from evolving data streams. They use an integration of open source products that are used to control the data stream and facilitate the harnessing of knowledge from the data stream. The authors highlight that the CluSandra algorithm exhibits the following characteristics: configurable, distributable, elastically scalable, highly available and reliable, and simpler to implement

III. CLUSTERING METHODOLOGIES

Clusters divide data into groups so that we can gather beneficial and meaningful information. To obtain meaningful information clusters should capture the natural structure of data. In some scenarios cluster analysis is only a useful starting point for other purposes, such as data summarization. Hence it is very critical how data is clustered.

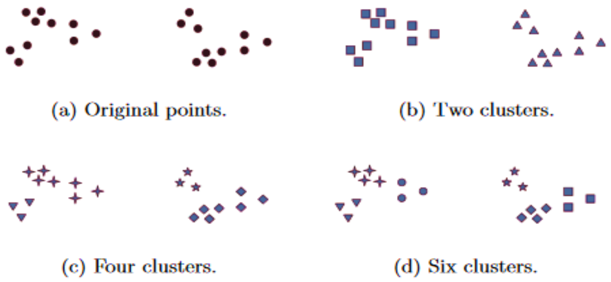

Figure 1: Different ways of clustering the same set of points

Figure 1 shows the different ways of dividing twenty points - 1(a) into 3 different clusters - 1(b), 1(c) and 1(d). The shapes of the points identify each cluster group. Each figure divides the data differently. Figures 1(b), 1(c) and 1(d) divide the data into two, four and six parts, respectively. The division of data may simply be an artifact of the human visual system. This figure illustrates that the definition of a cluster is imprecise and that the best definition depends on the nature of data and the desired results [15].

At a high-level Clustering algorithm are classified as Partition based, Hierarchical based, Density based, Grid based and Model based. An overview of different classifications of clustering algorithms is given below.

A. Partition based Clustering algorithms

All objects are initially considered as a single cluster. The objects are divided into partitions with each partition representing a cluster. Partitioning algorithms are K-means, K-medoids (PAM, CLARA, CLARANS, and FCM) and K-modes. All these techniques are based on the idea that a centre point can represent a cluster. The partition based algorithms work well for finding spherical-shaped clusters in small to medium-sized data points [14].

Advantages

- Relatively scalable and simple.
- Suitable for well separated datasets with compact spherical clusters

Disadvantages

- In high dimensional spaces, the concept of distance between points is ill-defined
- Pre-defined cluster count
- High sensitivity to initialization phase, noise and outliers

B. Hierarchical based Clustering algorithms

Hierarchical clustering can be done using the Agglomerative or Divisive technique depending on whether the hierarchical breakdown is by top-down (Agglomerative) method or bottoms-up (Divisive) method. In Agglomerative approach, initially one object is selected and successively merges (agglomerates) with its closest similar pair based on similarity criteria until all the data forms a desired cluster. The Divisive approach starts with a single cluster and divides the cluster into additional clusters down the hierarchy until the desired number of clusters are formed. BIRCH, CURE, ROCK, Chameleon, Echidna, Wards, SNN, GRIDCLUST, CACTUS are some of the hierarchical clustering algorithms.

Advantages

- Embedded flexibility regarding the level of granularity.
- Well suited for problems involving point linkages, e.g. taxonomy trees.

Disadvantages

- Inability to make corrections once the splitting/merging decision is made.
- Lack of interpretability regarding the cluster descriptors.
- Vagueness of termination criterion.
- Too expensive for high dimensional and massive datasets.
- Highly ineffective in high dimensional spaces.

C. Density based Clustering algorithms

Based on density, clusters are formed. Density based clusters are separated from each other by regions of low density objects. The low-density objects are noise or outliers. Various clustering algorithms form arbitrary shaped clusters such as DBSCAN, OPTICS, DBCLASD, GDBSCAN, DENCLU and SUBCLU. DBSCAN is one of the most commonly used clustering algorithms and the most cited in scientific literature. In 2014, this algorithm was awarded the test of time award (an award given to algorithms which have received substantial attention in theory and practice) at a leading data mining conference, KDD [4].

Advantages

- Discovery of arbitrary-shaped clusters with varying size
- Resistance to noise and outliers

Disadvantages

- High sensitivity to the setting of input parameters
- Poor cluster descriptors
- Unsuitable for high-dimensional datasets because of dimensionality phenomenon.

D. Grid based Clustering algorithms

In Grid based clustering algorithms, the data space is partitioned into cells to form a grid like structure. Then working on each cell multi-resolution clustering is performed. Since Grid algorithms perform the clustering on the grid versus the database they have much faster processing power when compared to other algorithms. Some grid based algorithms are STING, CLIQUE, Wave cluster, BANG, OptiGrid, MAFIA, ENCLUS, PROCLUS, and STIRR [10].

Advantages:

- Efficient for large multidimensional spatial databases
- Insensitive to outliers and the data input order

Disadvantages:

- Need to tune grid size and density threshold
- Cn have high mining costs

E. Model based Clustering algorithms

In model-based clustering algorithms, clusters are formed using models. An ideal fit between the model and data determines the cluster assignment. In the model-based clustering approach, the assumption is that the data is produced by a combination of probability distributions in which each module characterizes a different cluster. This algorithm works well if the data aligns with a model [6]. Algorithms such as EM, COBWEB, CLASSIT, SOM, and SLINK are well known Model based clustering algorithms.

Advantages:

- Since models are a comparison, models can abstract away from details to capture a general insight

Disadvantages:

- When generalized models are more complicated
- Which model to compare with is a big exploration

IV. RESULTS AND DISCUSSION

Throughout the literature paper review the one thing that stood out is that the authors are trying to find better clustering methods that would benefit in obtaining smart data from big data. Teams are trying to achieve this by using faster analytical approaches and turning weaknesses of individual algorithms into strengths by preprocessing or by combination of clustering methods. The authors feel a combination of clustering techniques would benefit the grouping of data since the best clustering methods are combined and between them they form new solutions that possess better results than their individual antecessors. For example [10] by using grid based algorithm for its speed and hierarchical algorithms for its flexibility the authors can perform clustering at different levels of resolutions and dynamically discover nested clusters. The authors [5] propose CluSandra, a clustering framework and algorithm not just for datasets but also for data streams. All the papers lead to the next steps of advanced algorithms, better data mining techniques, machine learning, cloud computing, and IOT.

Authors Osama [1], Priyanka [12], Richa & Jitendra [2] and Prakash &Aarohi [13] in their papers have provided a comparison on clustering algorithms. Results and analysis given below Table 1. The algorithms are chosen based on popularity, flexibly, ability to handle high-dimension data sets and applicability. The authors use software tools such as WEKA, LNKNet, Cluster and TreeView Packages for comparisons.

Table 1: Various Clustering Type comparison

Clustering Type	**Algorithm Name**	**Number of Clusters (k)**	**Cluster Instance/distribution**	**Number of Iterations**	**Time Taken to Build (sec)**
Partition Based	K-Means	2	643 (64%) 357 (36%)	5	0.03
Hierarchical Based	Birch	2	999 (100%) 1 (0%)		7.12
Density Based	DBScan	2	336 (99%) 4 (1%)		0.35
Model Based	SOM	4	12 (29%) 9 (22%) 15 (37%) 5 (12%)		0.33

After analyzing the results of their experiments, the authors conclude that k-means clustering algorithm is the simplest and fastest algorithm compared to the other algorithms. Other observations:

- Hierarchical clustering algorithm is more prone to noisy data
- Hierarchal algorithm takes more time than k-mean algorithm
- Density based algorithm takes relatively less time to build a cluster
- Density based clustering algorithm is not suitable for data having very huge variations in density
- SOM shows more accuracy in classifying objects into their appropriate clusters
- A general conclusion is partition based algorithms are used for smaller data sets and hierarchical for larger data sets

As a real-life example to better understand the usage of clustering algorithms we have used the Oil and Gas (O&G) industry. Clustering is helpful in discovering patterns from Oil and Gas E & P (Exploration and Production) data. In Clustering, both the dense and sparse regions of datasets are plotted and using the concept of distance metric or similarity metric clusters are formed. Since petroleum data is actual numbers used in statistical applications and pattern recognition, a large class of metrics exist. Based on requirements, metrics are defined. In a typical petroleum database, the number of attributes could be very large, while the size of an average transaction or its attribute value is small. Furthermore, basins having similar oilplay, for example, similar reservoir patterns belonging to a single local cluster, contribute to large scale spatial extents of reservoir attributes within the petroleum system and thus enhancing knowledge on reservoir presence at large scale [11]. In the E&P sector unplanned interruptions costs billions of dollars per year. To address this problem O&G companies need better data storage and analytic techniques of sensor data. Clustering is used in this sector for data mining.

V. CONCLUSION

With the dawn of Big data, clustering assists in grouping objects to analyze information, recognize patterns, and simplify the data. In this paper we have shown how clustering is performed, the different classifications, advantages and disadvantages of clustering types and several authors views on clustering. We have also shared experiment results on the different clustering methods. All algorithms have some defect in certain aspects hence we propose Cross-Clustering (CC). A partial clustering algorithm which combines the best of two or more well established clustering algorithms should serve the purpose. CC performs better than the other methods in: identification of outliers, estimating the correct number of clusters and real cluster membership. This method has been used in the medical field to identify disease subtypes and gene profiles to determine groups of genes with same behavior. This could also be used on non-biological datasets.

REFERENCES

[1] O.A. Abbas, "*Comparisons Between Data Clustering Algorithms"*, International Arab Journal of Information Technology (IAJIT), Vol.5, Issue.3 pp.321-325, 2008.

[2] R. Agrawal, and J. Agrawal, "*Analysis of Clustering Algorithm of Weka Tool on Air Pollution Dataset*", International Journal of Computer Applications, Vol.168, No. 13, 2017.

[3] S. Äyrämö, and T. Kärkkäinen, "*Introduction to partitioning-based clustering methods with a robust example*. Reports of the Department of Mathematical Information Technology", Software engineering and computational intelligence. University of Jyväskylä, Finland, Series C, 1/2006.

[4] M. Dave and H. Gianey, "*Different clustering algorithms for Big Data analytics: A review*", In System Modeling & Advancement in Research Trends (SMART), IEEE International Conference (pp. 328-333). 2016

[5] J.R. Fernandez and E.M. El-Sheikh, "*CluSandra: A framework and algorithm for data stream cluster analysis*", International Journal of Advanced Computer Science and Applications, Vol.2, No.11, pp. 87–99, 2011.

[6] V.K. Gujare, P. Malviya, "*Big Data Clustering Using Data Mining Technique*", International Journal of Scientific Research in Computer Science and Engineering, Vol.5, Issue.2, pp.9-13, 2017.

[7] M. Halkidi and I. Koutsopoulos, "*Online clustering of distributed streaming data using belief propagation techniques*", In Mobile Data Management (MDM), 2011 12th IEEE International Conference on (Vol. 1, pp. 216-225), 2011

[8] AR. PonPeriasamy, E. Thenmozhi, "*A Brief survey of Data Mining Techniques Applied to Agricultural Data*", International Journal of Computer Sciences and Engineering, Vol.5, Issue.4, 2017

[9] J.F. Laloux, N.A. Le-Khac, and M.T. Kechadi, "*Efficient distributed approach for density-based clustering"*, In Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE), 2011 20th IEEE International Workshops on (pp. 145-150). 2011

[10] W.K. Liao, Y. Liu, and A. Choudhary, "*A grid-based clustering algorithm using adaptive mesh refinement*", In 7th Workshop on Mining Scientific and Engineering Datasets of SIAM International Conference on Data Mining (Vol. 22, pp. 61-69), 2004.

[11] S.L. Nimmagadda and H. Dreher, "*Petro-data cluster mining-knowledge building analysis of complex petroleum systems*", In Industrial Technology, 2009. ICIT 2009. IEEE International Conference on (pp. 1-8), 2009.

[12] P. Sharma, "*Comparative Analysis of Various Clustering Algorithms Using WEKA*", International Research Journal of Engineering and Technology (IRJET), Vol.2, Issue.04, 2015.

[13] P. Singh, and A. Surya, "*Performance Analysis of clustering algorithms in data mining in WEKA*", International Journal of Advances in Engineering & Technology, Vol.6, Issue.6, pp.1866-1873, 2015.

[14] Ruchi Jayaswal, Jaimala Jha , Ravi Devesh , "*An Effective Method of Image Mining using K-Medoid Clustering Technique*", International Journal of Computer Sciences and Engineering, Vol.5, Issue.6, pp.206-214, 2017.

[15] P.N. Tan, M. Steinbach and V. Kumar, "*Data mining cluster analysis: basic concepts and algorithms*", "Introduction to data mining", Pearson Education, India. pp. 487–569, 2013.

[16] I. Triguero, J. Maillo, J.Luengo, S. García and F. Herrera, "*From Big data to Smart Data with the K-Nearest Neighbours algorithm*", In Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData), 2016 IEEE International Conference on (pp. 859-864), 2016.