Performance Analysis and Evaluation of Clustering Algorithms

Sanskruti Patel, Atul Patel

Abstract*: *The Today's digital world, data generation is growing at a rapid rate, almost doubling every two years. The extensive growth in the digital devices generates and consumes enormous data. In the field of Artificial Intelligence, machine learning provides a paradigm to recognize hid-den patterns in the data to perform useful inference using those patterns that have been learned. Clustering or cluster analysis is one of the most essential and important unsupervised learning technique. Clustering is a technique of natural grouping of data objects which are unlabeled and it forms these grouping in such a way that data objects belonging to one cluster are not similar to the objects belonging to another cluster. In this paper, different clustering approaches and techniques used in unsupervised learning are discussed. Also, four major clustering algorithms namely k-means, EM, hierarchical and make density based are applied on different datasets and their performance is analyzed by using certain parameters.

Index Terms*: *Unsupervised learning, Clustering algorithms, K-means clustering, Hierarchical clustering, Density-based Clustering.

I. INTRODUCTION

In today's digital world, data generation is growing at a rapid rate, almost doubling every two years. It has been predicted that, every second 1.7 megabytes of new information will be created for every human being by the year 2020. The extensive growth in the digital devices generate and consume enormous data. These data sets are in form of structured, semi-structured or unstructured. However, such large data sets are not useful without analytic power. To extract meaningful information from these massive and varying data sets, some data analytics process should be applied. Machine Learning is a field of Artificial Intelligence (AI) that consists a set of algorithms to recognize patterns in data to perform useful implications. It consists set of algorithms used and executed by computer systems for automatically perform certain tasks without explicit directives. For these, it relies on patterns and inference mechanism. For the extraction of data, various techniques are available for customization for the specific set of information. Among of these techniques, clustering is a technique of natural grouping of data objects which are unlabelled and it forms these grouping in such a way that data objects belonging to one cluster are not similar to the objects belonging to another cluster[1]. Clustering or cluster analysis is one of the most essential and important unsupervised learning technique. It covers the three well-known categories of cluster analysis namely partition, hierarchical and density-based clustering. The algorithms considered for an experiment in this study are k-means clustering algorithm; EM algorithm make density based clustering algorithm and hierarchical clustering algorithm. All the mentioned algorithms are explained and analysed based on the certain evaluation parameters. These parameters are the number of clusters created, incorrectly clustered instances, time taken to build the model and other parameters like number of instances in the dataset and type of the data set.

II. AN OVERVIEW OF MACHINE LEARNING TECHNIQUES

Machine learning algorithms own a capability of self-learning and improvisation. Machine learning algorithms are useful to discover the hidden pattern from the massive and heterogeneous datasets. They normally categorized into three categories: Supervised Learning, Unsupervised Learning and Reinforcement Learning as shown in following figure 1.

Supervised learning algorithms aim at categorizing data from prior information[2]. These algorithms apply on a set of input variables P and an output variable Q. The prediction of output for unseen data is completed through a mapping between P and Q. Supervised algorithms are applied on dataset that contains labelled data with input and desired output [3]. It consist two approaches: classification and regression. Naïve Byes, Support Vector Machine, Decision Tree, Multilayer perceptron neural network are some the well-known algorithms categorized under supervised learning. Unsupervised learning algorithms infers from datasets consisting of input data without labelled output data. It consist two approaches: clustering and association rule mining. The data set assigned to an algorithm does not consist pre-defined labels like supervised learning. Cluster analysis using k-means clustering, hierarchical clustering and density based clustering are most widely algorithms follows unsupervised approach. Reinforcement learning is a type of learning where software agents must to take actions in a way that will maximize some notion of cumulative reward. These actions may affect situations and their actions in the future [4].

The algorithms falls under unsupervised learning relies on mathematical model based on data set that contains data without output labels. They learns from test data that has not been labelled, classified or categorized. Unsupervised learning basically infers from dataset that consists input data without label output data and it manly focuses on identifying commonalities in the data and for each new data, it matches based on absence or

Revised Manuscript Received on December 22, 2018.

Sanskruti Patel, Faculty of Computer Science and Applications, CHARUSAT, Changa, Gujarat, India
Atul Patel, Faculty of Computer Science and Applications, CHARUSAT, Changa, Gujarat, India

presence of commonalities. Therefore, the dataset passed to unsupervised learning algorithm contains only input variables with no associated output variable and it is expected that the algorithms are self-capable to discover and exemplify the interesting structure in the data. Normally unsupervised algorithms are used to find patterns form in the data normally by grouping the data points. There are normally two approaches are available for unsupervised learning problems: clustering and association. Clustering normally groups the data and finds the hidden patterns from the data. Association mainly discovers rules for describing large portions of the data.

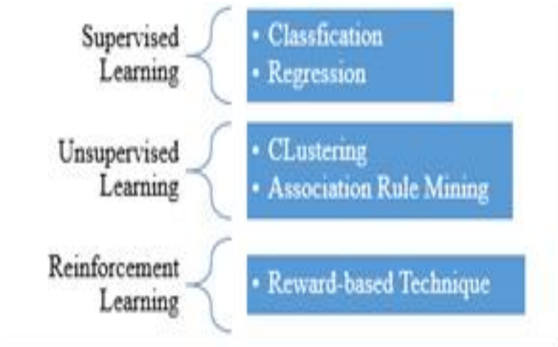

Fig 1. Types of Machine Learning Algorithms

III. CLUSTERING: AN UNSUPERVISED LEARNING TECHNIQUE

One of the most common techniques in unsupervised learning is clustering or cluster analysis. It is one of the significant technique for pattern recognition and machine learning [5]. It is used for empirical data analysis to uncover hidden patterns or groupings in data. Cluster analysis normally group the similar data into one group and other are into different group. Clustering is a widely used technique for data analysis and pattern discovery, which is used in many fields [6].

As clustering falls in an unsupervised learning mechanism, it forms clusters based on set of patterns (data) in a way that members of one cluster are similar according to a predefined criterion [7]. Each cluster consists of data, which are homogeneous, and they are heterogeneous compared to data of other cluster[8]. A partition between the data elements are chosen in such a way that they minimize some measure of dissimilarity between members of the same cluster [9]. On pool of multidimensional data, clustering technique uses some similarity measure normally Euclidian distance and it is also termed as a difficult problem as the clusters may vary in their sizes and shape [10]. The most common and frequently used cluster algorithms are k-means clustering, k-medoids, hierarchical clustering, hidden Markov model, fuzzy c-means clustering, subtractive clustering and Gaussian mixture models [11]. Moreover, these systems are broadly divided into three categories namely hierarchical, partitioning and density based clustering [12] as per shown in following figure 2.

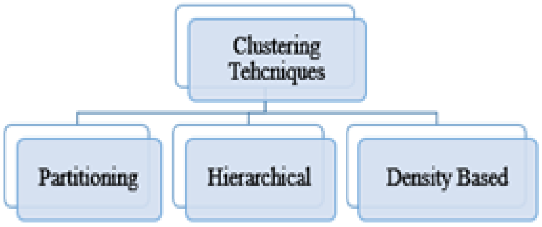

Fig 2. Categories of Clustering Techniques

I. Partitioning Clustering

Partitioning clustering are clustering methods used to classify data points into multiple groups based on their similarity. To generate number of clusters, algorithms required to specify it initially. It decomposes a dataset into a set of separate clusters. Partitioning clustering algorithms are iterative algorithms and they divide the data set into a specified number of clusters. K-means is one of the most widely used partitioning algorithm for cluster analysis [13]. The K-Means algorithm divides n object into K clusters in such a way that clusters have relatively high similarity in the cluster and, relatively low similarity between clusters [14] [15]. To form a cluster centre, mean value is to be calculated from each cluster. The calculation of similarity is done by mean value of the cluster objects. To calculate the distance, Euclidean distance method is used. The following is the procedure of K-means algorithm.

1. For generation of data members to be clustered, place K points into the space. For initial group centroids, these points are considered.
2. Assign each data member to the group that has the closest centroid.
3. When all objects have been assigned, recalculate the positions of the K centroids.

Repeat Steps 2 until the centroids no longer move. This produces a separation of the data members into groups from which the metric to be minimized can be calculated. Usually, the K-mean algorithm criterion function adopts square error criterion, defined as:

$$z=\Sigma\Sigma\|(j)-cj\|2$$

In the formula, $\|(j)-cj\|2$ is the distance between the data point $xi(j)$ and the cluster centroid cj. The following figure 2 shows how the data points are clustered by an algorithm. The K-means algorithm has the following main advantages [16].

- It is widely used clustering algorithm and easy to implement.
- It is even suitable for very large data set.

Some of the drawback of K-means algorithms are as follows [17].

- The users need to specify the number of clusters initially.
- It is highly dependent on data.
- Unable to identify outliers and noise.

II. Hierarchical Clustering

Hierarchical clustering is mainly focuses on building of hierarchy of clusters, i.e. cluster tree and it is represented in a dendrogram [18]. It is either merging smaller clusters into larger clusters or splitting larger clusters into smaller ones. A clustering of the data items is obtained through cutting a dendrogram at a desired level [19]. A cluster tree is defined as "a tree showing a sequence of clustering with each clustering being a partition of the data set” [20]. Following are the general procedures for performing hierarchical clustering [21].

1. Assign each item to a cluster and therefore of items are X, there will be X clusters.
2. The distance between the clusters are same as the distance between the items contained by cluster.
3. Then after, find the closest pair of clusters and merge them into a single cluster so one cluster is removed.
4. Compute the distance between the exiting clusters and the newly formed cluster.
5. Repeat the above two steps till all items are grouped into K number of clusters.

Hierarchical clustering is further classified into two categories: Agglomerative and Divisive.

Agglomerative algorithms: These algorithms merges the smaller clusters into larger ones. It is also known as bottom up approach [22][23]. It starts with 1 point. It is then after add two or more appropriate clusters recursively. Finally, it stops when k number of clusters is achieved [13]. It is too slow for large data sets as its complexity is O (n^3) [9].

Divisive algorithms: Divisive algorithms performs splitting of larger clusters into smaller ones. It is also called a top down approach. It starts with a big cluster and then recursively divides into smaller clusters [25]. It stops when k number of clusters achieved [13]. Its complexity is O (2n) which is not adequate [9].

Some of the advantages of using Hierarchical Clustering algorithms are as follows [17].

- Initialization is not required.
- Performs well even on noisy data.
- Clustering is similar to that perceived by humans.
- Simple and good for small data set.

Some the drawbacks observed are as follows [6,9].

- Computationally expensive and not suitable for large datasets.
- Difficulty handling different sized clusters and convex shapes
- Static by nature as patterns assigned to one cluster cannot move to another cluster.

III. Density-based Clustering

In density-based clustering, the regions with higher density is considered in a data space as compared to the regions with lower density. The clusters discovers in density-based clustering are of arbitrary shapes and they are adequate for handling noise. Here, the term density is mainly focuses on the minimum numbers of points within a certain distance of each other [12]. It is wildly used in data, which has noise, and when there are outliers present in the data. At the time of grouping the clusters, distance measurement and a minimum number of points are considered, known as Eps and Min Points [24]. Outliers are formed by considering data points falls in low density regions. The minimum distance between two points are defined by Eps. When the distance between two data points is either less or equal to Eps, these points are considered as neighbours [19]. Some of the advantages of using make density based algorithms are as follows [9].

- It is mainly useful with noisy data and when there are outliers present on the data set.
- Its result is closer to K-means algorithm.
- It returns density and distribution.

Some of the drawbacks of using make density based algorithms are as follows [9].

- Density measures affects by sampling.
- It is sensitive to clustering parameters used.

IV. EXPERIMENTS AND RESULTS

I. Dataset description

There are five datasets obtained from UCI data repository to carry out an experiment of how clustering algorithms works and analysis of their performance. The following table 1 describes the dataset name, number of attributes in each dataset, number of instances, missing value in dataset and dataset type. The dataset taken for experiment are possessing different characteristics and falls in different field of applications.

Dataset Name	**Dataset Characteristics**	**Missing Value**	**No. of Attributes**	**No, of Instances**
Absenteeism at work	Multivariate, Time-Series	N/A	21	740
Diabetes 130	Multivariate	N/A	22	7195
Mice Protein Expression	Multivariate	Yes	82	1080

Table1. Description of the Datasets used in the Experiment

II. Results and Discussion

The experiment is carried out using Weka 3 Toolkit. It consists of a collection of machine learning algorithms for data mining tasks. It is an open source software in Java, freely available under General Public License (GNU) agreement and offers a very good user interface for the different tasks like pre-processing, classification, clustering etc. In the experiment, four clustering algorithms namely k-means, hierarchal, make density based and EM. Moreover, there are three datasets chosen for conducting an experiment. For performance analysis, the parameters like incorrectly clustered instances, no. of iterations and time taken to build model are considered. The following table 2 shows the result of experiment conducted with dataset and other parameters. Based on the results, it has been observed that k-means clustering algorithm performs well in all three datasets. Its accuracy is compare to more and the time taken to build the model is also adequate. Make density based algorithm also works well with good accuracy and considerable time taken to build the model. It has been also observed that, hierarchal clustering and EM clustering algorithm provided less accuracy and consumed more time. Therefore, it has been perceived from results that k-means clustering algorithms performs well on all three different types of datasets.

Data set	Clustering Algorithm	No. of Clusters and its Distribution	Incorrectly Clustered Instances	No. of Iterations	Time taken to build model (sec)
Absenteeism at work	K-means Clustering	0 - 394 (53%) 1 - 346 (47%)	78.1081 %	11	0.12
	Hierarchal clustering	0 - 739 (100%) 1 – 1 (0%)	79.7297 %	--	0.57
	Make Density Based Clustering	0 - 407 (55%) 1 - 333 (45%)	78.3784 %	11	0.05
	EM	0 - 238 (32%) 1 - 502 (68%)	79.1892 %	2	0.06
Anuran Calls (MFCCs)	K-means Clustering	0 - 2466 (34%) 1 - 4729 (66%)	36.0806 %	15	0.13
	Hierarchal clustering	0 - 6653 (92%) 1 - 542 (8%)	44.1279	--	450.67
	Make Density Based Clustering	0 - 2836 (39%) 1 - 4359 (61%)	36.2891 %	15	0.12
	EM	0 - 3728 (52%) 1 - 3467 (48%)	36.6644 %	3	0.91
Mice Protein Expression	K-means Clustering	0 - 569 (53%) 1 - 511 (47%)	73.6111 %	7	0.06
	Hierarchal clustering	0 - 1080 (100%)	86.1111 %	--	4.55
	Make Density Based Clustering	0 - 534 (49%) 1 - 546 (51%)	73.7963 %	7	0.07
	EM	0 - 505 (47%) 1 - 575 (53%)	74.0741 %	29	0.65

Table 2. Performance Analysis of Clustering Algorithms

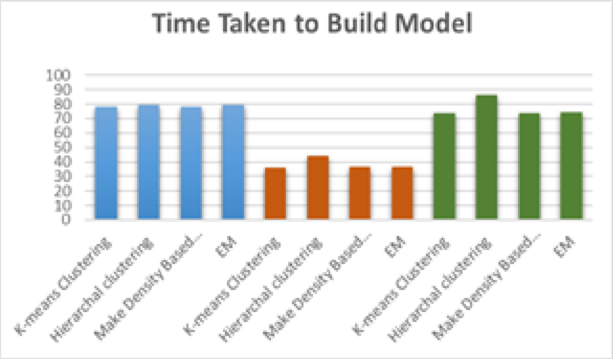

Fig 3. Time taken to build model by different algorithms

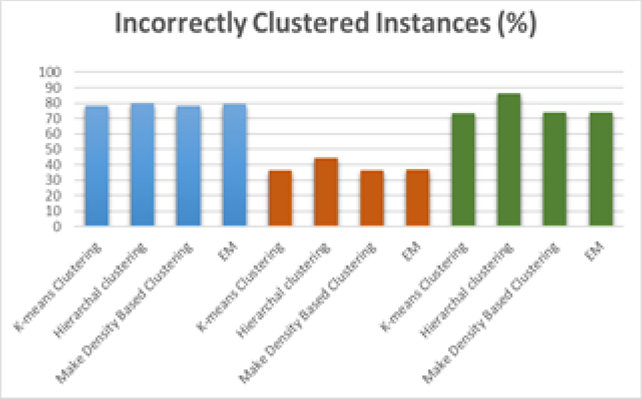

Fig 4. Incorrectly clustered instances by different algorithms

V. CONCLUSION

Huge data is accumulated at every second with the integration of digital technology in various fields including medical, banking, education, retail and many more. These data are in heterogeneous form and it is required to apply ample data analytics process to extract meaningful information from this massive datasets. The paper comprises the study of various clustering algorithms that can be applied to discover the hidden patterns from the dataset. It also covers a comparative analysis of the performance of k-means, hierarchal, make density based and EM algorithms experimented on different data sets. In future, the experiment will also carry out on huge and vague datasets with inclusion of more clustering algorithms.

REFERENCES

1. Proceedings of the Second International Conference on Soft Computing for Problem Solving (SocProS 2012), December 28-30, 2012", Springer Nature America, Inc, 2014
2. A. Singh, N. Thakur and A. Sharma, "A review of supervised machine learning algorithms," 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, 2016, pp. 1310-1315
3. S.B. Kotsiantis, “Supervised Machine Learning: A Review of Classification Techniques”, Informatica 31 (2007) 249-268
4. Ayon Dey, “Machine Learning Algorithms: A Review”, International Journal of Computer Science and Information Technologies (IJCSIT), Vol. 7 (3), Pages. 1174-1179, 2016

5. G. Hamerly and C. Elkan, “Alternatives to the K-means Algorithm that Find Better Clusterings”, In Proceedings of the ACM Conference on Information and Knowledge Management (CIKM-2002), pp. 600-607, 2002
6. Omran, Mahamed & Engelbrecht, Andries & Salman, Ayed. (2007), “An overview of clustering methods”, Intell. Data Anal.. 11. 583-605. 10.3233/IDA-2007-11602
7. Bhoopender Singh, Gaurav Dubey, “A comparative analysis of different data mining using WEKA”, International Journal of Innovative Research and Studies, ISSN: 2319-9725, Volume 2, Issue 5, Page 380-391, May 2013
8. Dr.Naveeta Mehta, Shilpa Dang, "A Review of Clustering Techniques in various Applications for Effective Data Mining", International Journal of Research in IT & Management, ISSN 2231-4334,Volume 1, Issue 2, Page 50-66, June 2011
9. Peerzada Hamid Ahmad, Dr. Shilpa Dang, Performance Evaluation of Clustering Algorithm Using Different Datasets, Journal of Information Engineering and Applications, Vol.5, No.1, 2015, pp. 39-47
10. Jain, R. Duin and J. Mao, “Statistical Pattern Recognition: A Review”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no.1, pp. 4-37, 2000.
11. https://in.mathworks.com/discovery/machine-learning.html
12. Prakash Singh, Aarohi Surya, “Performance Analysis of Clustering Algorithms In Data Mining In Weka”, International Journal of Advances in Engineering & Technology, January 2015, Vol. 7, Issue 6, pp. 1866-1873
13. Wang L., Bo L., Jiao L. (2006, “A Modified K-Means Clustering with a Density-Sensitive Distance Metric”, In: Wang GY., Peters J.F., Skowron A., Yao Y. (eds) Rough Sets and Knowledge Technology. RSKT 2006. Lecture Notes in Computer Science, vol 4062. Springer, Berlin, Heidelberg
14. A. Wosiak, A. Zamecznik and K. Niewiadomska-Jarosik, "Supervised and unsupervised machine learning for improved identification of intrauterine growth restriction types," 2016 Federated Conference on Computer Science and Information Systems (FedCSIS), Gdansk, 2016, pp. 323-329
15. Sunila Godara and Ritu Yadav, “Performance analysis of clustering algorithms for character recognition using weka tool”, International Journal of Advanced Computer and Mathematical Sciences, ISSN 2230-9624. Vol 4, Issue 1, 2013, pp119-123
16. R.H. Turi, “Clustering-Based Colour Image Segmentation”, PhD Thesis. Monash University, Australia, 2001
17. E. Davies, “Machine Vision: Theory, Algorithms, Practicalities”, Academic Press, 2nd Edition, 1997
18. Dong W., Ren J., Zhang D. (2011), “Hierarchical K-Means Clustering Algorithm Based on Silhouette and Entropy”, In: Deng H., Miao D., Lei J., Wang F.L. (eds) Artificial Intelligence and Computational Intelligence. AICI 2011. Lecture Notes in Computer Science, vol 7002. Springer, Berlin, Heidelberg
19. Zhao Y., Karypis G., “Evaluation of hierarchical clustering algorithms for document datasets”, the eleventh international conference on Information and knowledge management,2002, pp. 515-524.
20. Y. Leung, J. Zhang and Z. Xu, “Clustering by Space-Space Filtering”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no.12, pp. 1396-1410, 2000
21. Jain A.K., Murty M.N., Flynn P.J., “Data clustering: A Review”, ACM Computing Surveys, Vol. 31, No. 3, September 1999
22. Aarya Vardhan Reddy, Paakaala Sai Saran Macha, Kumara Saketh Mudigonda, “Evaluation of Clustering Algorithms on Absenteeism at Work Dataset”, IJSRD - International Journal for Scientific Research & Development| Vol. 6, Issue 06, 2018 ISSN (online): 2321-0613 , pp. 337-342
23. T. Sajana, C. M. Sheela Rani , K. V. Narayana, “A Survey on Clustering Techniques for Big Data Mining”, Indian Journal of Science and Technology, Vol 9(3), DOI: 10.17485/ijst/2016/v9i3/75971, January 2016
24. Glory H. Shah, C. K. Bhensdadia, Amit P. Ganatra, An Empirical Evaluation of Density-Based Clustering Techniques, International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-1, March 2012, pp. 216-223
25. Amol Bhagat, Nilesh Kshirsagar, Priti Khodke, Kiran Dongre, Sadique Ali, "Penalty Parameter Selection for Hierarchical Data Stream Clustering", Procedia Computer Science, 2016