A Comparative Study on Unsupervised Clustering Machine Learning Algorithms

Leo Vattoly
Department of Computer Science
Central University of Kerala
Tejaswini Hills, Periye, Kerala, India- 671320
leo.vattoly@icloud.com

***Abstract*—Clustering is the most useful and important data mining methods for knowledge discovery in large dataset. It is the process of organizing and dividing the large data sets according to their similarity [1]. Clustering is also known as an unsupervised machine learning technique because there is no prior data supplied by the practitioner. Clustering algorithm makes the data to be labelled using the statistical technique. It has a major role in unsupervised machine learning, mostly used in research areas, which help the practitioner to find the hidden patterns and relate the data entity. Clustering divides the data into multiple regions, every region object shows the similarities in its region and dissimilarities with another region object. There exist several wide varieties of algorithms to cluster the data. The main objective of this paper is to provide a general comparative study on clustering algorithms partition-based *(k-Means)* and Density-Based *(DBSCAN)*.**

***Index Terms*—unsupervised Learning, Clustering, Partitioning algorithm, k-Means, Density Based algorithm, DBSCAN**

I. Introduction

Machine learning has made a great leap in all the field, it conquering all the R&D areas and enhancing the capability of the machine to an extended level. It took the machine to the next dimension, act like a human being or replace the human effort. Nowadays machines can predict or suggest according to the data trained by the practitioner, this practice is called supervised learning. While training the machine the practitioner facing some challenges, such as making Knowledge Discovery in Databases (KDD) from a set of raw data. It can be avoided by the unsupervised machine learning mechanism. In unsupervised learning technique, the machine learns from hidden patterns, which obtained from a set of unlabelled datasets. The term cluster has a great role in unsupervised learning, especially to catch the hidden patterns from the datasets. We are formulating a comparative study between two different approaches of clustering, partitioning method Density-based method [2] [3].

The paper is formulated as, section 2 provides the methodology adapted for the study. In the next section related work is specified, it includes the different approaches in clustering algorithms and limitations of the algorithms. The fourth section covers the result obtained by the study takes place. Conclusion and the future works are included in the fifth section.

II. methodology

We are comparing the partition and Density-based algorithm based on entropy with Iris dataset.

The overall procedure for studying algorithms are added following, first collecting the dataset from different data repositories such as Kaggle(Online dataset competition platforms cum community). Next, we are clustering the dataset using the k-means and DBSCAN. From the clustered model, we are obtaining the Entropy from the clustered dataset, which used for the cross-validation.

III. Related Work

A. Partitioning Method

*k-means*is the most popular and simple clustering algorithm. It has been discovered by several researchers across different methods, most remarkably Lloyd (1957, 1982), Forgy (1965), Friedman and Rubin (1967), and McQueen (1967) [4].

This algorithm helps the practitioner to partition the data into a specified number of clusters within a group. Suppose k-means algorithm works on a data set D, contains n data objects in Euclidean space, partitioning method will distribute the data set D into k different cluster where k $\leq$ n [5]. Initially algorithm select k different centroids, which can be either systematically or randomly. That makes the n objects in to k number of clusters by default. Then the algorithm take place an iteration with two different steps to achieve a stable centroid value. The first step is allocating all points to the closest centroid and, again measure the centroid of each cluster [6].

Algorithm: k-Means.

k: The number of clusters.
D: A data set consist of n objects.

```
Select k objects as the initial centroids from D.
repeat
    lifted each object to the immediate centroid.
    update the centroid value of each cluster.
until no change in centroid
```

The k-means algorithm works on a set of objects D with a dimension of d , where the partitioning method distribute the

Figure 1 and figure 2 are taken from the slides for the book, *Introduction to Data Mining*, Tan,Kumar,Steinbach, 2006.

objects into k different clusters. Consider $C_1, C_2, \ldots, C_k$ as the cluster classes, $C_i \subset$ D and $C_i \cap c_j = \emptyset$, $1 \leq$ i and j $\leq$ k. We can measure the quality of a cluster C_i using the sum of squared error between all objects in C_i and the centroid c_i [5].

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} Euclideandistance(p, c_i)^2 \qquad (1)$$

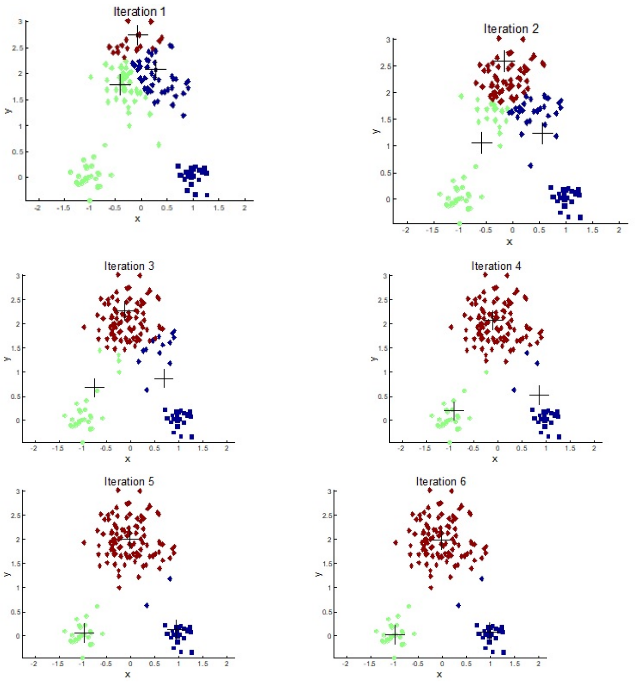

Fig. 1. Centroid changing in every iteration (+ sign indicates the centroid)

Limitations of *k-means*: The partitioning method can be misunderstood the object and cluster it in wrongly by differing in size, densities, non-globular shapes. Although false clustering may occur when the data contains outliers. Its very difficult to cluster the arbitrary shaped objects using k-means [7].

B. Density-Based Method

Density Based Spatial Clustering of Application with Noise (DBSCAN) algorithm is the most convenient algorithm in Density based clustering paradigm, which can detect arbitrary shaped clusters.This algorithm introduced in 1996 [8] [5] [9].

The core idea of the DBSCAN algorithm is to form clusters have high enough density, which can be specified by the practitioner using the following parameters Eps_i (The radius value of the cluster) and $MinPts_i$ (The smallest number of objects inside the Eps_i region)

The following terms help us to define the DBSCAN algorithm.

1) **Neighbourhood**: Any dataset D, which contains an object m, the neighbourhood of Eps is a set of spherical collection all objects formed within, a circle with m, Eps is the radius of the area, denoted as:

$$NEps(m) = n \in \text{ D, distance(m,n)} \leq \text{Eps} \qquad (2)$$

2) **Core Object**: For any object in the DB, which having less than MinPts inside the region of Eps. Then the object is a core object.
3) **Noise**: Let $C_1, \ldots, C_k$ are the clusters in the dataset D wrt. Eps_i and $MinPts_i$, i=1,. . . ,k. Then we can decide a point which exist in D but not to C_i.
4) **Density-reachable**: A point p is density attainable from a point q wrt. Eps and MinPts. if there is a group of points $p_1, ..., p_n, p_1 = q, p_n = p$ such that p_i+1 is precisely density-reachable from p_i.
5) **Density-connected**: A point p is density correlated to a point q wrt. Eps and MinPts if there is a point o such that both, p and q are density-reachable from o wrt. Eps and MinPts.
6) **cluster**: A cluster C_i wrt. Eps and MinPts is a non-empty subgroup of D fulfilling the consecutive conditions:
   a) each p,q :if p resides to Ci and q density reachable from p wrt.Eps and MinPts, then q belongs to C
   b) for every p,q $\in$ C: p is density associated to q wrt. Eps and Minpts

To detect a cluster, DBSCAN opening with an arbitrary point p and fetch all points density-reachable from p wrt. Eps and MinPts. If p is a focus point, this action yields a cluster C wrt. Eps and MinPts, and all the points of NEps(q) reside to it. If p is a adjoin point, no points are density-reachable from p and DBSCAN inspect the later point of the data set [10] [11].

Algorithm: DBSCAN

```
1 Choose an arbitrary point p.
2 Fetch complete points density-reachable from p wrt. Eps
  and MinPts
3 If p is a root point, a cluster is develop.
4 If p is abut point, no points are density attainable from p
  and DBSCAN inspect the later point of the database.
5 Extend the action until all the points have been refined.
```

Limitations of *DBSCAN*: It can perfectly cluster the objects which have a large difference in density, but it will struggle when the objects are having similar density.The parameters(Eps & MinPts) of the algorithm are highly sensitive, hence it may include errors to the clusters.

IV. Experimental Result

The lower the value of entropy indicates good clustering. For the same data sets, the DBSCAN and K-means algorithms were obtained and the clusters were generated. Table 1 indicated values show that the DBSCAN has a low value of entropy. This indicates good clustering.

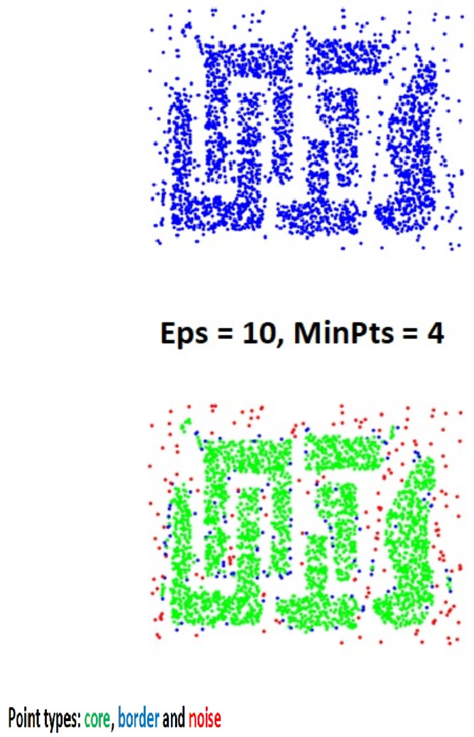

Fig. 2. : It shows the differences once the DBSCAN algorithm applied to a dataset. We can clearly find the clustering of arbitrary region of objects)

TABLE I
EXPERIMENTAL RESULT

Sl.No	**Result Based on Entropy**		
	Algorithm	***Entropy***	
1	k- Means	1.584	
2	DBSCAN	0.918	

V. CONCLUSION AND FUTURE SCOPE

The paper discussed the most commonly used unsupervised clustering algorithms. The aim of the paper was to prepare a comparative study between the Partition and Density based clustering algorithms. The paper shows that density-based algorithm shows good clustering while taking entropy as quality measuring technique.

REFERENCES

[1] D. Birant and A. Kut, “St-dbscan: An algorithm for clustering spatial–temporal data,” *Data & Knowledge Engineering*, vol. 60, no. 1, pp. 208–221, 2007.

[2] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

[3] A. Mucherino, P. Papajorgji, and P. M. Pardalos, *Data mining in agriculture*, vol. 34. Springer Science & Business Media, 2009.

[4] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, S. Y. Philip, *et al.*, “Top 10 algorithms in data mining,” *Knowledge and information systems*, vol. 14, no. 1, pp. 1–37, 2008.

[5] J. Han, J. Pei, and M. Kamber, *Data mining: concepts and techniques*. Elsevier, 2011.

[6] Y. Shi and D. Olson, *Introduction to business data mining*. McGraw-Hill/Irwin, 2007.

[7] P.-N. Tan, M. Steinbach, and V. Kumar, “introduction to data mining. ed ke-1,” 2006.

[8] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, *et al.*, “A density-based algorithm for discovering clusters in large spatial databases with noise.,” in *Kdd*, vol. 96, pp. 226–231, 1996.

[9] E. Yasser, M. Ismail, and M. Farouk, “An efficient density based clustering algorithm for large databases,” in *Proc. 2004 16th IEEE Int. Conf. Tools with Artificial Intelligence (ICTAI 2004)*, 2004.

[10] X. Fu, Y. Wang, Y. Ge, P. Chen, and S. Teng, “Research and application of dbscan algorithm based on hadoop platform,” in *Joint International Conference on Pervasive Computing and the Networked World*, pp. 73–87, Springer, 2013.

[11] M. Chen, X. Gao, and H. Li, “Parallel dbscan with priority r-tree,” in *2010 2nd IEEE International Conference on Information Management and Engineering*, pp. 508–511, IEEE, 2010.