A Comparative Study on Unsupervised Clustering Machine Learning Algorithms

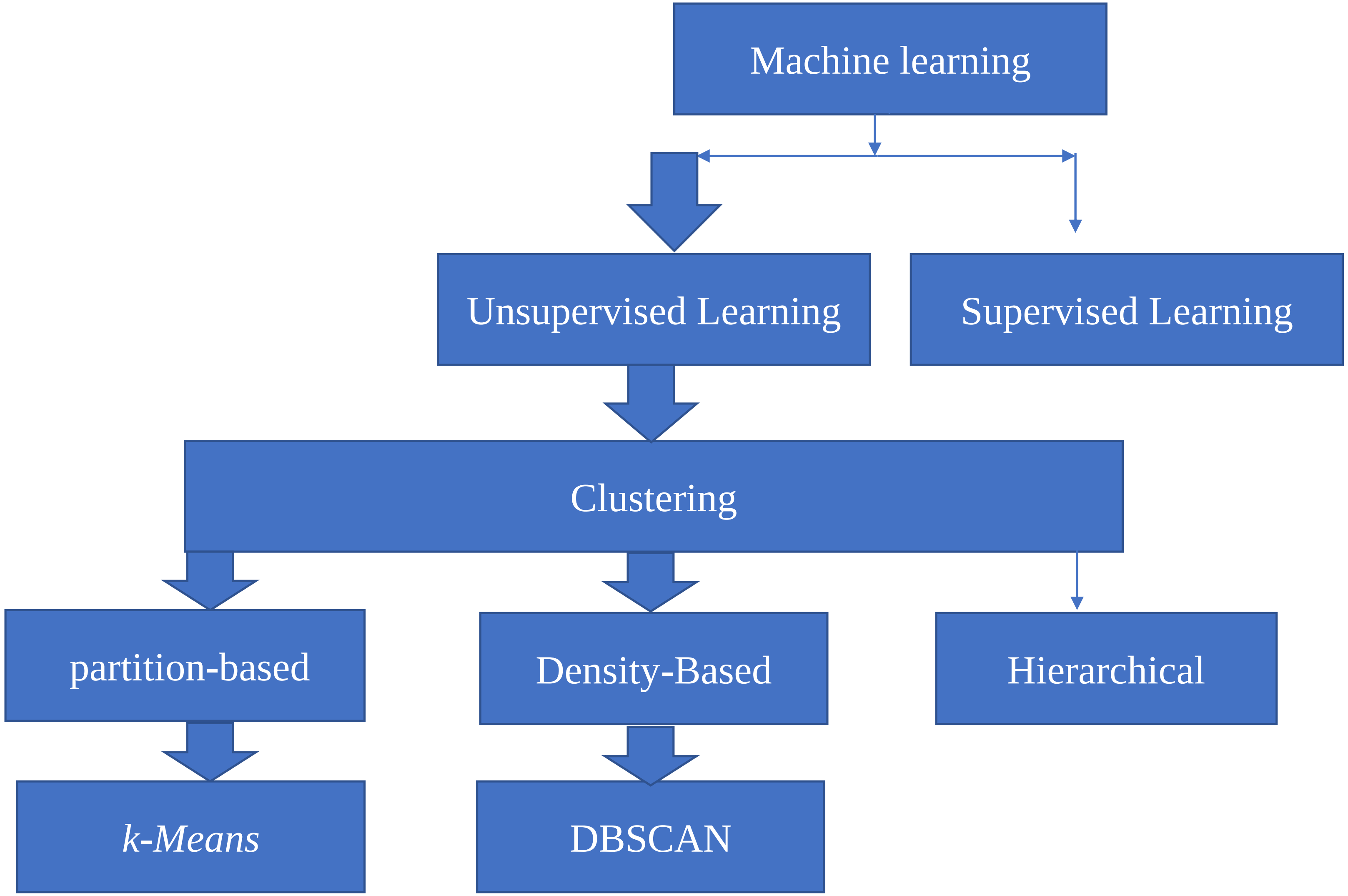
Leo Vattoly

Department of Computer Science

Central University of Kerala

Email: leo.vattoly@icloud.com

Introduction



What is machine learning?

Machine learning (ML) is the scientific study of algorithms and statistical models that computer systems use to perform a specific task without using explicit instructions, relying on patterns and inference instead.

What is Unsupervised learning ?

Machine learning from hidden patterns, which obtained from a set of unlabeled dataset.(Learning in the absence of prior data)

What is Clustering ?

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

Partition Based

- *k*-Means

Density Based

- *DBSCAN*

Partitioning Based Algorithm
k-Means

K means algorithm cluster the n number of data object into k different groups. In this algorithm practitioner has to share the value for k.

Algorithm

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

```
1 Select k objects as the initial centroids from D.
2 repeat
        3 (re)assign each object to the nearest centroid.
        4 update the centroid value of each cluster.
5 until no change in 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.
- Its very difficult to cluster the arbitrary shaped objects using k-means

Density Based Algorithm
DBSCAN

Density Based Spatial Clustering of Application with Noise (**DBSCAN**), which can detect arbitrary shaped clusters. In this algorithm practitioner has to share following parameters

- **Epsi** (The radius value of the cluster)
- **MinPtsi** (The smallest number of objects inside the Epsi region)

Algorithm

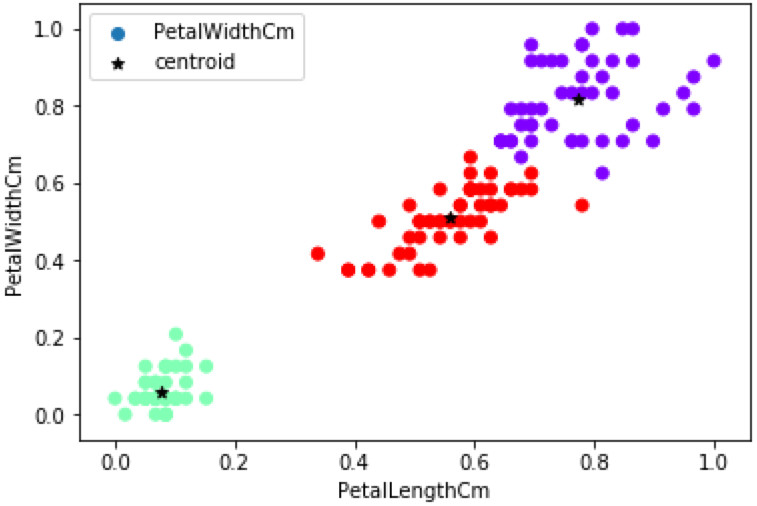
1. Select an arbitrary point p.
2. Retrieve all points density-reachable from p wrt. Eps and MinPts
3. If p is a core point, a cluster is formed.
4. If p is border point, no points are density reachable from p and DBSCAN visits the next point of the database.
5. Continue the process until all the points have been processed.

Limitations of DBSCAN:

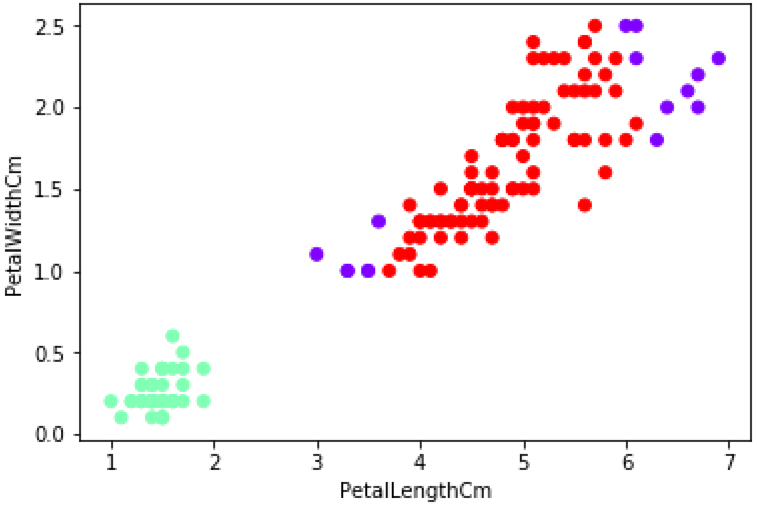
- It will struggle when the objects are having similar density.
- The parameters(Eps & MinPts) of the algorithm are highly sensitive, hence it may include errors to the clusters.

Experimental Result

We are comparing both Partition-based and Density-based algorithms using the IRIS dataset . Entropy as quality measuring factor.



K- Means output
K=3
Entropy = 1.58



DBSCAN output
eps=0.8
MinPtsi=50
Entropy = 0.918

Conclusion

- 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 having good clustering while taking entropy as quality measuring technique.

References

- 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.

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

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

- 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.

- Y. Shi and D. Olson, Introduction to business data mining. McGrawHill/Irwin, 2007.

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

- 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.

- 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.

- 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.

- 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.

THANK YOU