

DBSCAN Clustering

Tasawar Siddiquy
Electronic Engineering
Hochschule Hamm-Lippstadt
Lippstadt, Germany
tasawar.siddiquy@hshl.stud.de

Abstract—Clustering algorithms have been used in various applications. Clustering methods are very efficient to draw out details of spatial data from different applications. It can be used to extract useful patterns from complicated data sources. But all the clustering algorithms are not efficient to use for every purpose. Generally in the case of an arbitrary figure in databases including noise and outliers, not all the clustering algorithms work accurately whereas DBSCAN (Density-Based Spatial Clustering of Applications with Noise) can handle these kinds of data sets. However, DBSCAN clustering also has some flaws which have been enhanced by researchers and some different variants also have been introduced. Although We will mainly focus on one of the most well-recognized DBSCAN clustering as well as how it can be implemented for extracting important pieces of information from databases in this research paper.

Index Terms—Machine learning, clustering, DBSCAN clustering, algorithm

I. INTRODUCTION

Machine learning is a technique of creating models that can analyze and learn from datasets. We have done extensive improvements in the Computer sector by using machine learning methods[5]. Machine learning methods becoming very favored progressively and the implementation results of machine learning are also very impressive. We constructed many different kinds of methods and appliances to make our life easier and more efficient. Human beings always discovering and inventing things to fulfill tasks with perfection and with less effort. That's why researchers are continuously developing new methods and patterns. The system which has been configured with machine learning algorithms can habituate its behavior to real-time inputs like a Human. The system can also anticipate the future using analyzed data and inconsistent information[4]. As an example, suppose several events happening in a system following some protocols or patterns. The result of the system depends on the events or the datasets. So if we create a pattern from the datasets and predict some future consequences, think that the pattern is a sort of behavior of the system as well as teach the machine how to behave and handle such events without certain inputs that are called machine learning. There are three kinds of machine learning: supervised learning, unsupervised learning (for an example DBSCAN clustering), and reinforcement learning. Machine learning is a subtype of Artificial intelligence and deep learning is a subset of machine learning[fig 1]. Although they have a similar purpose but the working processes are not the same.

This was a short introduction to machine learning but in this paper one of the most well-known machine learning algorithm will be introduced which is the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) clustering.

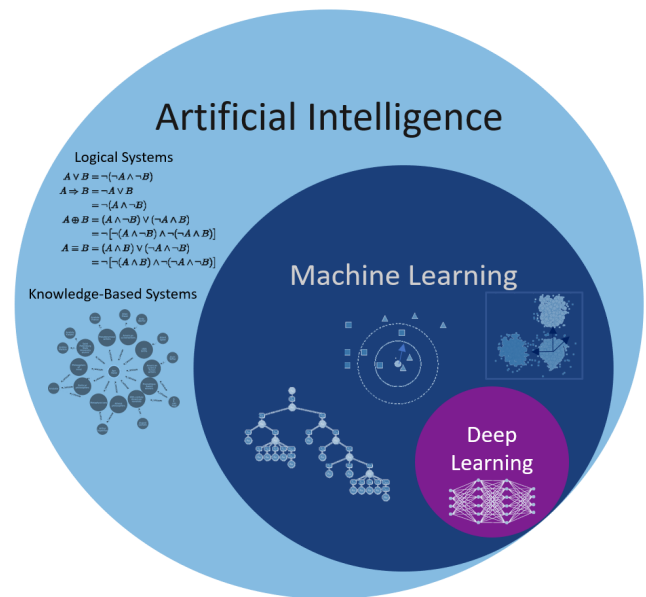


Fig. 1. AI(ML(DL))

<https://data-science-blog.com/blog/2018/05/14/machine-learning-vs-deep-learning-wo-liegt-der-unterschied/>

II. DBSCAN CLUSTERING

Clustering means the grouping of objects or dividing each instance into several groups. Clustering algorithms are widely used. . Clustering algorithms are unsupervised and they have a wide range of applications in Data analysis, pattern recognition, machine learning, image processing, market research, data mining, spam identifying, and other areas.

DBSCAN is one of the most prominent clustering algorithms [7][8]. DBSCAN clustering depends on the density of the regions. The efficiency of the DBSCAN clustering algorithm performs better if the clusters are dense enough [1]. The objective of this kind of unsupervised machine learning method is to extract hidden patterns and cluster similar data

[2]. DBSCAN or Density-Based Spatial Clustering of Applications with Noise is an uncomplicated and effective algorithm that is capable of recognizing any number of clusters, of any shape. Noise means which data is purposeless from the main sequence of data. values in a data pack. DBSCAN does not need any initial statement about the number of predicted clusters. It can solve easily non-convex problems that K-means clustering fails to execute [4]. DBSCAN clustering is slightly slower than agglomerative clustering and K-means, despite DBSCAN regardless scales to moderately large datasets [3]. The drawbacks in nearly all of the traditional clustering algorithms are heightened computational complicatedness and they can not scale well the bigger size datasets [6].

III. THE ALGORITHM

The primary idea of density-based clustering algorithms is to construct a cluster that is dense enough and divided by a low-density region [10]. DBSCAN clustering was developed to cluster data of random shapes including the noise in spatial and non-spatial high dimensional datasets [6]. Multiple applications need the management of spatial data means the data which is related to space [9]. The Eps(epsilon) in Figure 2 is a defined radius that is estimated by calculating the data number inside the Eps-neighborhood of a point. The main theory of the DBSCAN clustering algorithm is that each point in the cluster must have at least a Minpts(minimum points) number of points together with itself inside its Eps-neighborhood. DBSCAN clustering starts with selecting an arbitrary point and checks if it includes neighboring points less than the Minpts number of points, then it is marked as noise temporarily or else it constructs a cluster. After that, the points inside the Eps-neighborhood of the selected points are counted to the cluster and the cluster begins to extend. The user can specify two parameters in DBSCAN: Eps and Mints [10].

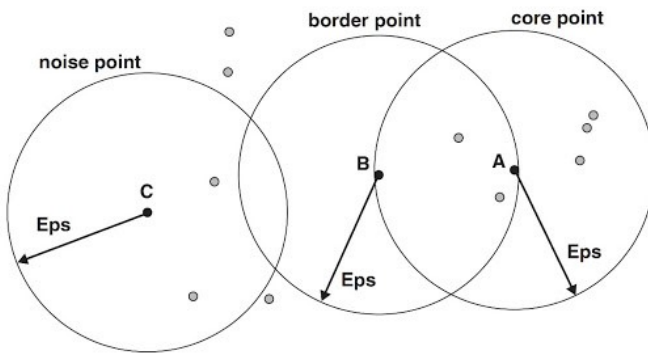


Fig. 2. Core point, border point, noise point

<https://www.analyticsvidhya.com/blog/2021/06/understand-the-dbscan-clustering-algorithm/>

$$1) NEps(A)=B \in |dist(A, B) \leq Eps)$$

NEps indicates to the Eps-neighborhood of a point A, dist(A,

B) is a distance function for A and B points. NEps(A) gathers points with a distance equal to or less than Eps from A.

- 2) $B \in NEps(A)$
- 3) $|NEps(A)| \geq Minpts$

B point is directly-reachable from point A if it follows the Condition number 1 and 2. Condition number 3 is for the core point in Figure 2. So as long as a point's neighborhood includes at least Minpts, we can label it a core point. If Minpts = 5, then we can say that A is the core point. Border point means that is not a core point but it falls into the eps-neighborhood of a core point.

So in Figure 2, we can call B a border point and B is density-reachable from point A. Using condition numbers 1 and 2 the algorithm will search for more core points. If the conditions are matched it will expand the clustering step by step. According to this in Figure 3 we can see that the clustering area has been extended because conditions 1 and 2 were matched.

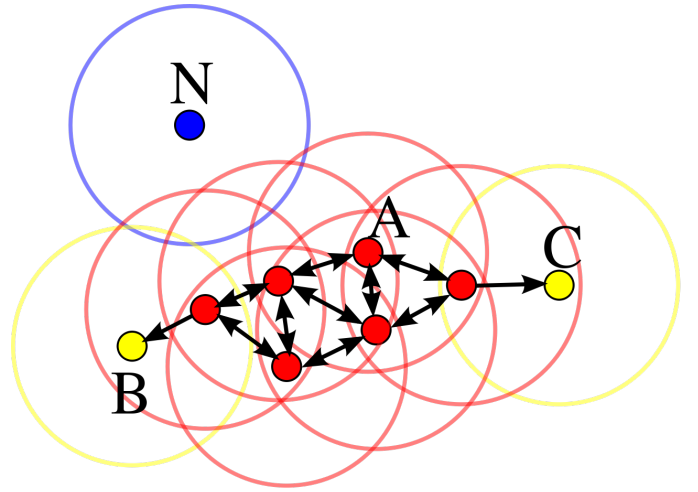


Fig. 3. Density-connected

<https://en.wikipedia.org/wiki/DBSCAN>

In Figure, 3 points C and B are the border point and they are density-reachable as well as we can say density-connected. In Figures 2 and 3 points, C and N are not clustered because they did not meet any conditions So these are the noise points [6][7][9][10]. The clustering result can differ if the Eps value is not correct. So the user should check the dataset and determine an appropriate Eps value[10].

In Figure 4 we can observe how the data has been separated into clusters by using the DBSCAN algorithm [8]. Several small clusters can be analyzed to learn important information from the dataset. The noise data also has been isolated from the dataset.

Many types of research have been done to improve and enhance the performance of the DBSCAN clustering algorithm. Some modified clustering methods also have been introduced based on the DBSCAN clustering technique [6][10][12][13][14].

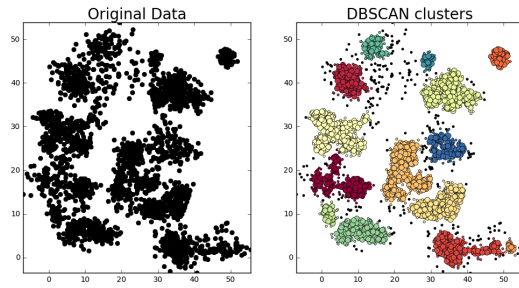


Fig. 4. Clustered Data

<https://github.com/chriswernst/dbscan-python>

IV. APPLICATIONS OF DBSCAN CLUSTERING

Nowadays clustering algorithms have been implemented in various applications. Several applications of the DBSCAN method are the following:

- **X-ray crystallography:** A real-world application that locates all the atoms or particles on the inside of a crystal, which retrieve a big amount of data. Then the different kinds and quantity of atoms in the data which has been fetched from the crystal should be discovered and categorized, this process can be done by using the DBSCAN algorithm [12].
- **Satellited images:** Many images are being taken by satellite every day but these images are not categorized. The images must be organized or labeled to retrieve useful pieces of information. As an example to specify more information in digital maps, mountains or forests can be categorized using the DBSCAN technique [12].
- **Anomaly Detection in Temperature Data:** The applications that highlight design anomalies in the dataset, which is important in different cases like Healthcare, credit fraud, and also measuring the changes in temperature, which is also useful because of the global warming or ecological changes. So the unusual or abnormal design in the data can be detected by using the DBSCAN algorithm and later on analyzed to get control [12].
- **Spam Identifying:** presently Anti-spam mechanism is being used very widely. To identify spam more accurately, similar kinds of emails should be identified or clustered. So some new techniques based on DBSCAN clustering can be used to identify spam [8].
- **Suspicious financial transactions Identifying:** The particular data or information of a client as monthly deposit frequency and deposit amount, monthly money withdrawal amount, and monthly withdrawal frequency, these data can be clustered by the DBSCAN algorithm. The separated data can be examined or investigated to identify suspicious transactions [11].
- **Evaluating students learning status:** Evaluating students learning status: Flawed teaching effect is a common concern in the classic teaching process. So DBSCAN clustering method can be used to categorize students and develop a corresponding teaching technique according to

different classes of students. Firstly the characteristics of the students should be extracted as a dataset like self-control index, knowledge level, and career orientation. Furthermore, this method already has been verified [15].

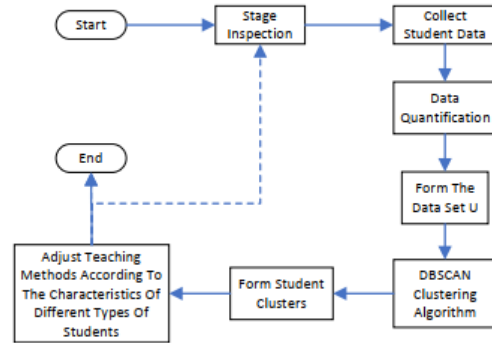


Fig. 5. Flow chart for algorithm framework [15]

In Figure 5, the following algorithm framework can be implemented: Select a specific stage for testing or complete the interest questionnaire as well as the Self-Control scale. Then Collect student data and quantify the three-dimensional data after that the student data set can be clustered by applying the DBSCAN algorithm. The final step is to adapt the teaching strategy according to the attributes of different categories of students [15].

- **New employee adaptation:** The DBSCAN clustering algorithm can be implemented to digitalize the process of adaptation of new employees in a company. In the first place financial costs are gathered into account and a web interface for questioning, and getting information has been applied [16].

V. ENHANCED DBSCAN CLUSTERING

DBSCAN clustering is a very well-operating machine learning method but it is not appropriate for all sorts of databases. Because the data attributes are too close to each other they will be clustered together and the result of clustering may not be accurate. Sometimes it also lacks performance such as the clustering time as well as the data sets containing various data points from time to time some important data could be clustered as a noise which also may result in wrong clustered data. And one of the main reasons behind these consequences is the parameters of DBSCAN clustering which the user has to observe from the data sets and specify before the clustering. But the data density is not the same in all the areas which are why the general value of Eps and Minpts is not suitable sometimes [10]. Therefore to prevent such problems many researchers have proposed so many enhanced DBSCAN techniques and also some of its variants. Of those modified techniques some of them were effective like the performance and accuracy of the result were improved compared to the traditional DBSCAN clustering. So we will discuss in detail some of these methods [6][10][12][13][14].

In 2013, Manisha et al. presented an enhanced algorithm that can detect and define the input parameters automatically based on the knowledge obtained from the database, and these manually defined parameters are one of the main drawbacks of the DBSCAN clustering algorithm. It is also very effective for large data sets. It identifies the cluster intuitively by accurately discovering the input parameters and is also able to cluster with differing densities. Using different Eps values it is feasible to find the clusters from varied densities continuously. For each value of Eps, the DBSCAN method is embraced to make sure that all the cluster's corresponding densities are clustered properly. After that, the clustered points are ignored which avoids labeling denser areas as one cluster. The testing result shows that the proposed technique can identify clusters of different densities with diverse arbitrary shapes and sizes from big amounts of datasets which include noise and outliers. So basically it needs only one input parameter and provides more acceptable output than the DBSCAN clustering algorithm [6].

VI. CONCLUSIONS

REFERENCES

- [1] A. Géron, *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: concepts, tools, and techniques to build intelligent systems*, Second edition. Sebastopol, CA: O'Reilly Media, Inc, 2019.
- [2] Sjardin, B., Massaron, L., Boschetti, A. (2016). *Large Scale Machine Learning with Python*. Packt Publishing.
- [3] A. C. Müller and S. Guido, *Introduction to machine learning with Python: a guide for data scientists*, First edition. Sebastopol, CA: O'Reilly Media, Inc, 2016.
- [4] G. Bonaccorso, *Machine Learning Algorithms Popular Algorithms for Data Science and Machine Learning*, 2nd Edition. Birmingham: Packt Publishing Ltd, 2018.
- [5] V. Lakshmanan, S. Robinson, M. Munn, and an O. M. C. Safari, *Machine Learning Design Patterns*. 2021.
- [6] K. Khan, S. U. Rehman, K. Aziz, S. Fong and S. Sarasvady, "DBSCAN: Past, present and future," *The Fifth International Conference on the Applications of Digital Information and Web Technologies (ICADIWT 2014)*, 2014, pp. 232-238, doi: 10.1109/ICADIWT.2014.6814687.
- [7] D. Deng, "DBSCAN Clustering Algorithm Based on Density," *2020 7th International Forum on Electrical Engineering and Automation (IFEEA)*, 2020, pp. 949-953, doi: 10.1109/IFEEA51475.2020.00199.
- [8] Wu Ying, Yang Kai and Zhang Jianzhong, "Using DBSCAN clustering algorithm in spam identifying," *2010 2nd International Conference on Education Technology and Computer*, 2010, pp. V1-398-V1-402, doi: 10.1109/ICETC.2010.5529221
- [9] J. Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD'96)*. AAAI Press, 226–231.
- [10] W. Wang, Y. Wu, C. Tang and M. Hor, "Adaptive density-based spatial clustering of applications with noise (DBSCAN) according to data," *2015 International Conference on Machine Learning and Cybernetics (ICMLC)*, 2015, pp. 445-451, doi: 10.1109/ICMLC.2015.7340962.
- [11] Y. Yang, B. Lian, L. Li, C. Chen and P. Li, "DBSCAN Clustering Algorithm Applied to Identify Suspicious Financial Transactions," *2014 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery*, 2014, pp. 60-65, doi: 10.1109/CyberC.2014.89.
- [12] J. P. Singh and P. A. Meshram, "Survey of density based clustering algorithms and its variants," *2017 International Conference on Inventive Computing and Informatics (ICICI)*, 2017, pp. 920-926, doi: 10.1109/ICICI.2017.8365272.
- [13] J. Wei and S. sun, "Commercial Activity Cluster Recognition with Modified DBSCAN Algorithm: A Case Study of Milan," *2019 IEEE International Smart Cities Conference (ISC2)*, 2019, pp. 228-234, doi: 10.1109/ISC246665.2019.9071776.
- [14] D. Jain, M. Singh and A. K. Sharma, "Performance enhancement of DBSCAN density based clustering algorithm in data mining," *2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)*, 2017, pp. 1559-1564, doi: 10.1109/ICECDS.2017.8389708.
- [15] H. Du, S. Chen, H. Niu and Y. Li, "Application of DBSCAN clustering algorithm in evaluating students' learning status," *2021 17th International Conference on Computational Intelligence and Security (CIS)*, 2021, pp. 372-376, doi: 10.1109/CIS54983.2021.00084.
- [16] M. A. Durova, E. A. Khodunov, A. S. Anikeeva, A. A. Mishin, A. N. Zein and S. V. Borisova, "Application of the dbscan Method to Solve the Problem of New Employees Adaptation," *2022 4th International Youth Conference on Radio Electronics, Electrical and Power Engineering (REEPE)*, 2022, pp. 1-6, doi: 10.1109/REEPE53907.2022.9731381.