Density-based clustering

written by Radu Galan

1. **Abstract**

While there is a continuous demand for better and better algorithms that can handle more and more data, there is a solution so good that has already been around for over 2 decades and we are still counting its usages. It has been first considered as an algorithm purposed for spatial data only since it came out when GPS and medical scan were providing heavy quantities of multidimensional data, but since then it’s been used in many other situations.

1. **Introduction**

\*

The purpose of this very paper is the analyze and clarify the density-based clustering algorithm by going through its creation, history, uses, and many more.

\*\*

This work has multiple motivations given the high demand in the spatial databases (and not exclusively) algorithms:

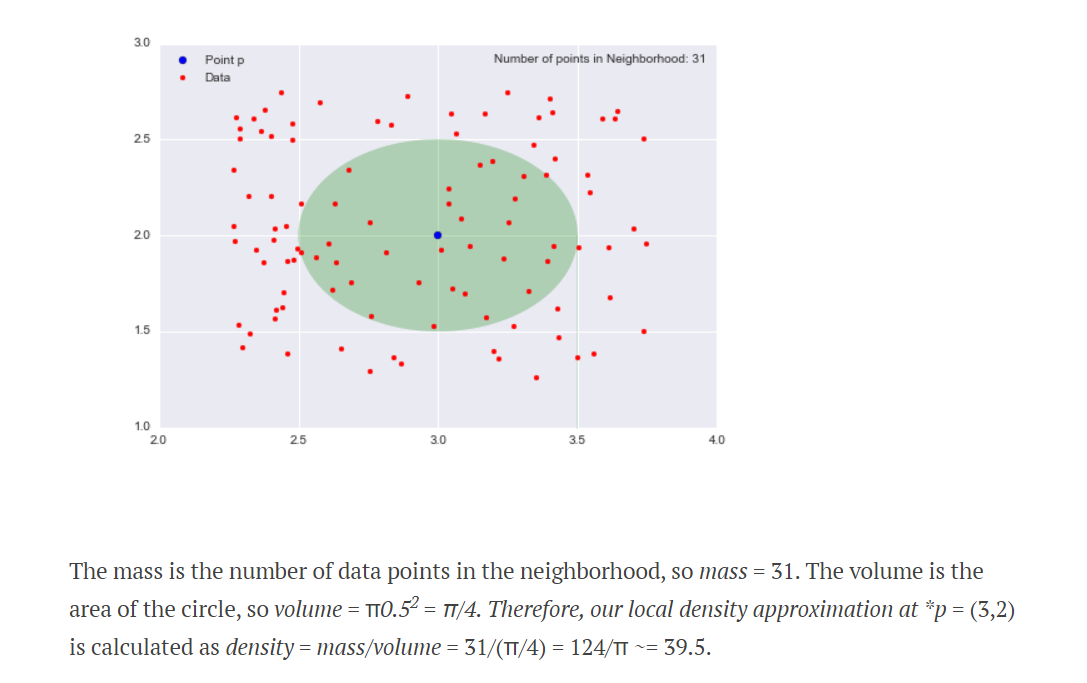
a. Usually there must be a minimal level of expertise in the given field, and this “requirement of domain knowledge” creates unnecessary additional steps in the process of data analysis. It would be much easier if a data scientist would not require expertise in a specific subfield in order to be able to analyze and make sense of data.

b. Most clustering methods work best for circular/spherical/.. clusters, but real life situations are not so regular as different compositions can be obtained: nonconvex, drawn-out, linear, elongated, and basically any possible imaginable shape (only restricted by the number of dimensions).

c. There is a serious need for faster algorithms since the sizes of databases are constantly growing. Usually creating one algorithm or solution for a data science problem doesn’t work efficiently for too long. Increasing the size of a database, even if the computational power is increased, is often a “deal-breaker” for most algorithms especially when clustering requires a polynomial complexity.

\*\*\*

This algorithm introduces the notion of density for data which can simply be explained by the physics formula:

In our situation the density can be calculated for each individual data entry. The volume will be the total space that consist the proximity of the center point (in our example it is a circle, but it can be a sphere/hypersphere/etc if the distance is constant) which implies the need of a definition for distance (by default the Euclidean distance can be used).

The mass will be equal to the number of points within the proximity of the center point. As we can observe the notion of weights for the entries could be very easily further introduced.

\*\*\*\*

We have presented an abstract for the method, some motivations that founded it, and already gave a very short intuition to the algorithm. In the following chapters we will dive into the subject step by step with the “3.Large picture”

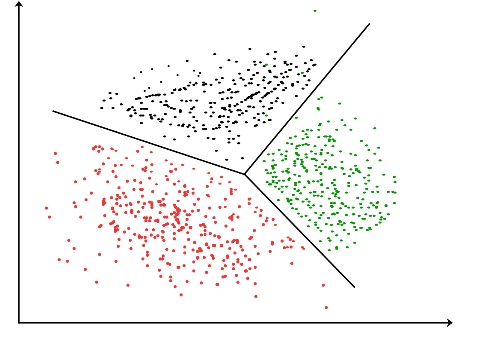
1. **Large picture**

\*

Density Based Clustering (DBSCAN) is a method designed and implemented in 1996 (as we will further see. This is an example of a learning algorithm that uses the unsupervised learning method of clustering.

Like any other learning algorithm from the data science subfield it works with data gathered from different sources (from sensors to other algorithms) with the sole purpose of extracting a certain information (usually as one or more formulas/rules that describe a certain situation or answer a particular question).

Unsupervised learning techniques in general are methods that can work with raw data and no labelling (categorizing each individual data entry as being part of a certain class) is required.

Clustering is an unsupervised method that works with the data and attempts to create classes based on the proximity (which is often described by Euclidean distance) of data entries to one another. The purpose is to create groups of those data entries where all the elements are as similar as possible.

\*\*

There are multiple methods of clustering and they can be divided in up to 7 categories (based on: (Sabhia Firdaus 2015)):

a. partitional (single level partition takes advantage of a centroid in order to best position a given number of clusters – K-means)

b. hierarchical (creates a tree of clusters based on data using operations like merging two clusters into one or dividing one cluster into two – BIRCH, ROCK)

c. density-based (introduces the notion of density and is able to approximate clusters with any form – DBSCAN)

d. grid-based (uses objects space between data instead of the data itself to divide classes – STING, Wave Cluster)

e. model-based (attempts learning cluster similarities based on probability distribution – Expectation-Maximization, Conceptual Clustering)

f. clustering high-dimensional data (CLIQUE, PROCLUS)

g. constraint based cluster analysis (the distance between two point is not as relevant as a set of given constraints)

Although a less rigorous classification would be simplified to just the first three: partitional, hierarchical, density-based. Those happen to be the most frequent used and it is very interesting that until 1996 when the first article of DBSCAN was published this very class was unknown to scientist and now over 24 years later it is still implemented in popular libraries, used, and even improved.

1. **The beginning**

\*

The algorithm was first proposed in [Martin Ester](https://en.wikipedia.org/wiki/Martin_Ester) (knowledge discovery and data mining with a MS degree in philosophy of science) , [Hans-Peter Kriegel](https://en.wikipedia.org/wiki/Hans-Peter_Kriegel) (spatial database systems and expert systems), Jörg Sander (database systems, query processing, similarity search,etc. who was a full professor), and Xiaowei Xu (data mining and high-dimensional indexing). All of them come from the Institute of Computer Science at the University of Munich and together have proposed a revolutionary algorithm called DBSCAN (density-based) in 1996. (Martin Ester, A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise 1996)

At the time only partitioning and hierarchical algorithms were used and this work was revolutionary for a reason. Although it was first considered to be used for spatial data bases only it became meaningful and efficient and almost any other situation of unsupervised learning dataset. Their first intuition was based on the idea that “we might want to discover classes of houses along some river”. They wanted versatility in clustering because algorithms could not properly categorize certain figures and also algorithms were not efficient enough for the size of their data.

\*\*

As we already explained the intuition for the algorithm in “2.Introduction” we will try and dive into more technical details. One main characteristic of the algorithm is that it does not require the number of clusters before-hand (this can be either an advantage or a disadvantage). In order the detect the local density it will need two main parameters:

1. Ɛ- the radius of the proximity of the datapoint
2. Minimum points- which represents the threshold for creating a cluster out of similar points

The most important distinction that DBSCAN makes between points is related to their density:

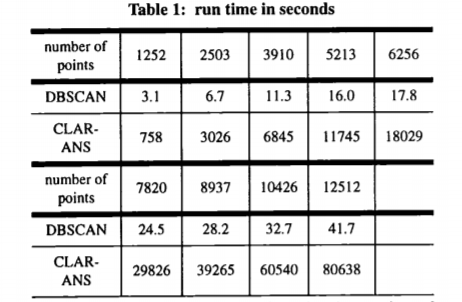
1. Core points (if it has in proximity the minimum of points with similar density)
2. Border points (if it is not a Core point but is in proximity of a Core point)
3. Outliers (if it is neither a Core point nor a Border point)

As we can clearly observe a very different characteristic has already been revealed. We now have for one of the very first times a clustering algorithm that also focuses on detecting outliers.

The main algorithm that could be used to solve this problem is CLARANS (Clustering Large Applications based on RANdomized Search) which explores partitioning algorithms for KDD in spatial databases (improved k-medoid method).

Apart from all the unique features it brings to the table the method is actually extremely efficient (and efficient enough even for our times) as it was over 100 times better than the next best of the time (CLARANS – (Raymond T. Ng 1994) ) and worked better in most situations.

The only similar work on this topic would be: (Prentice Hall. Kaufman L. 1988), but uses nonoverlapping cells with relatively high frequency and histograms. However, it is not a practical algorithm since it would have enormous runtime and space, but even so it is very dependent on the size of those cells.

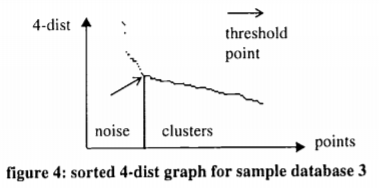
An algorithm called CLARANS (Clustering Large Applications based on RANdomized Search) is introduced which could be a potential opponent for this algorithm because there is no other that attempts the (knowledge discovery in databases (KDD)) problem.

The opponent algorithm does not work properly though because it splits data clusters if they are too big, there is no notion of outliers implemented, and the runtime is considerably higher.

\*\*\*\*

In the first paper (Martin Ester, A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise 1996) they create several definitions, iteratively creating all three classes that will represent the main returned types of labels: core points, border points, outliers. The basis for this are the concepts introduced: density-connected, density-reachable, directly density-reachable, neighborhood of an object.

The algorithm itself algorithm finds clusters by taking each individual point (possibly multiple times) and calculate its density. After that the concepts mentioned above will be applied in order to label the point as core, border, or outlier. Clusters will be formed based on an initial point that will initialize a new class and any point from its neighborhood (any point in range of another core point from this class) will be member of the same class. Clusters may be merged if the densities are similar.

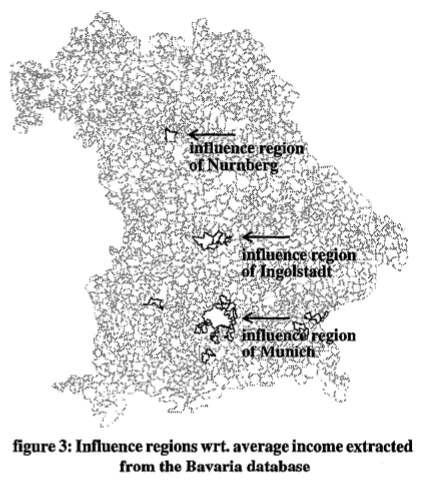
An important part is actually choosing the parameters. “We begin by mapping each point to the distance from its k-th nearest neighbor. When sorting the points of the database in descending order of their k-dist values, the graph of this function gives some hints concerning the density distribution in the database. We call this graph the sorted k-dist graph. need to find this threshold point where we have the needed values for the parameters.”

Although it is not yet optimized from a computational point of view, the results have been widely appreciated and the method itself created a new category of clustering algorithms all together and even received awards.

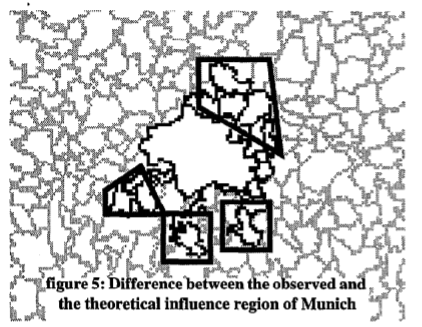
1. **Extensions**

They continued research with a few more papers: 1997, 1998, 2007 and even 2017 a few years after the algorithm have been awarded the Test of Time.

\*

In 1997: (Martin Ester, Density-Connected Sets and their Application for Trend Detection 1997) the same team of researchers came out with a so called ‘generalised’ version of the same algorithm: GDBSCAN. Although it didn’t really take off it was meant to extend the idea by creating a similar algorithm that would be able to deal with both spatial and non-spatial data and use them both. The motivation is ‘Trend detection’ for example finding rates of unemployment, moving trends and influences of certain regions.

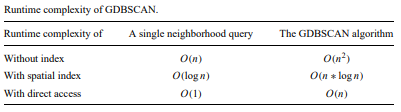
These spatial trends have been defines as “a pattern of systematic change of one or several non-spatial attributes in 2D or 3D space”.

Given enough parameteres it was actually able to properly predict the region of influence of multiple cities (for example Munich as can be seen in figure 5)

The motivation is quite high as economic geographers (that actually contributed to the study) have plenty of usages for this information).

\*\*

In the 1998 (Martin Ester, Density-Based Clustering in Spatial Databases 1998) paper they propose a list of multiple applications for the newly discovered GDBSCAN. Here they propose another version of the Generalized algorithm GDBSCAN, and then continue by providing four applications of the created algorithm: “2D points (astronomy), 3D points (biology), 5D points (earth science) and 2D polygons (geography)”.

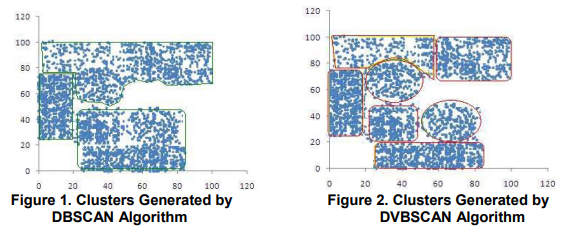


\*\*\*

Although there was another paper from 2017 (Martin Ester, DBSCAN Revisited, Revisited:Why and How You Should (Still) Use DBSCAN 2017) where they countered some criticism, there is not point for us to cover it now.

\*\*\*\*

There is also a very interesting extension of this algorithm that uses varied density and will only cluster together points of similar density (Prasad and Sarmah 2011) . Although it does need additional parameters it will be able to further create clusters based on local density. A proper comparison can be seen in the image below.



1. **Conclusion**

Density Based Clustering (DBSCAN) is a method designed and implemented in 1996 and was not awarded multiple prizes for nothing. It brings a whole new technique to research, a technique that proves itself to be incredibly optimal and efficient. This is certainly a technique that has shaped the image of clustering and could be used as an example of thinking outside the box anytime. The notion of density from physics has been used countless times since and it proves, from my point of view, how important interdisciplinary vision can be in coming with revolutionary ideas.

# 7.

# Bibliography

Martin Ester, Hans-Peter Kriegel, Jiirg Sander, Xiaowei Xu. 1996. "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise." *KDD-96 Proceedings* 226-231.

Martin Ester, Hans-Peter Kriegel, Jiirg Sander, Xiaowei Xu. 2017. "DBSCAN Revisited, Revisited:Why and How You Should (Still) Use DBSCAN." *ACM Transactions on Database Systems, Vol. 42, No. 3.*

Martin Ester, Hans-Peter Kriegel, Jiirg Sander, Xiaowei Xu. 1998. "Density-Based Clustering in Spatial Databases." *Data Mining and Knowledge Discovery 2,.*

Martin Ester, Hans-Peter Kriegel, Jiirg Sander, Xiaowei Xu. 1997. "Density-Connected Sets and their Application for Trend Detection." *KDD-97 Proceedings* 10-15.

Prasad, Rabinder Kumar, and Rosy Sarmah. 2011. "Variable density spatial data clustering." *2nd International Conference on Computer and Communication Technology (ICCCT-2011).*

Prentice Hall. Kaufman L., and Rousseeuw RJ. 1988. "Algorithms for Clustering Data." In *Data: an Introduction to Cluster Analysis. John Wiley & Sons*, by Jain Anil K. 1988.

Raymond T. Ng, Jiawei Han. 1994. "Efficient and Effective Clustering Methods for Spatial Data Mining." *Proceedings of the 20th VLDB Conference, Santiago Chile.*

Sabhia Firdaus, Md. Ashraf Uddin. 2015. "A Survey on Clustering Algorithms and Complexity." *IJCSI International Journal of Computer Science Issues, Volume 12, Issue 2, March 2015 .*