Ministry of Higher Education and Scientific Research

University of Babylon

College of Education for Pure Sciences

Mathematics department

Density-Based Spatial Clustering of Applications with Noise (DBSCAN)

A research submitted to the Council of the University of Babylon, College of Pure Sciences, Department of Mathematics, from the requirements of obtaining a bachelor's degree

By

Zahraa Kazem Khalif

Supervised by

Amera Abdul Wohid Funjan

2023 A.D **1444 A.H**

بِسْمِ اللَّـهِ الرَّحْمَـٰنِ الرَّحِيمِ

وَأَنْزَلَ اللَّهُ عَلَيْكَ الْكِتَابَ وَالْحِكْمَةَ وَعَلَّمَكَ مَا لَمْ تَكُنْ تَعْلَمُ ۚ وَكَانَ فَضْلُ اللَّهِ عَلَيْكَ عَظِيمًا

صدق الله العظيم

سورة النساء / آية (١١٣)

الاهداء

بِسۡمِ ٱللَّهِ ٱلرَّحۡمَٰنِ ٱلرَّحِيمِ

اللهم صل على محمد وال محمد وعجل فرجهم

ليت شعري اين استقرت بك النوى بل اي ارض تقلك او ثرى , ابرضوى او غيرها ام ذي طوى عزيز علي ان ارى الخلق ولا ارى ولا اسمع لك حسيسا ولا نجوى ,عزيز علي ان تحيط بك البلوى ولا ينالك مني ضجيج ولا شكوى ,,,

بنفسي انت من مغيب لم يخل منا بنفسي انت من نازح ما نزح عنا بنفسي انت من اثيل مجد لا يجاري بنفسي انت من تلاد نعم لا تضاهي ((الى متى احار فيك يا مولاي والى متى واي خطاب اصف فيك واي نجوى))

عندما تشتاق روحك وتنعم عينيك بطلته المباركة وبصوته الشجي وبصدره الحنون وبيديه الحانيتين وتفرح قلب امامك المنتظر حين يسمع منك صدق كلامك واخلاص قلبك ونيتك الصافية لرؤيته ومناجاته لذاته هو لا لشئ اخر ابدا ولا لاجل حل لمشاكلك او قضاء لحاجتك وانما شوق لسيدك فلنجعل انا املنا تخط اجمل العبارات الاهداء والتحايا الى صاحب العصر والزمان (ع) روحي وارواحكم له الفداء

اي نور انت حل بصلاتي وصيامي

اي نور انت حل بصباحي ومسائي

اي نور انت حل بحياتي وكياني

اي نور قد تجل منك يا كل مرامي

الشكر والتقدير

الحمد لله رب العالمين والصلاة والسلام على أشرف الأنبياء والمرسلين سيدنا محمَّد وعلى آله وصحبه ومن تبعهم بإحسان إلى يوم الدين، وبعد ..

فإني أشكر الله تعالى على فضله حيث أتاح لي إنجاز هذا العمل بفضله، فله الحمد أولاً وآخرًا.

اتقدم بالشكر والتقدير الى اهلي ثم إلى قمري الذي لا يغيب وشمسي التي لا ينقطع دفؤها أبداً إلى أغلى وأعز مخلوق عندي صاحب العقل الواعي والقلب الكبير والوجود المؤثر في حياتي رمز الكفاح وعنوان النجاح والدي المناضل الى الشمس التي أنارت دربي وأسعدت قلبي ودفأتني بحنانها إلى بسمة الحياة وأمل المستقبل وإشراقة النور أمي الحبيبة إلى الكواكب المشرقة والنجوم المتألقة والنخيل الباسقة إخواني وأخواتي إلى من رسموا لنا الطريق وساعدونا في اجتاز كل ضيق أساتذتنا الأفاضل إلى كل رفيق وصديق تمنى لي النجاح ودعاء لي كل من تمنى لي الخير والنجاح ، عائلتي وأصدقائي وصديقاتي وزملائي وزميلاتي والشكر الكبير لأعظم شخصين أبي أمي شكراً لكم بحجم السماء إليهم جميعاً أتقدم لهم بالشكر الجزيل راجياً من الله الإطالة بأعمارهم وأن يبارك فيهم ويحفظهم ربّي بعينه التي لا تنام كان يوم ليس كسائر الأيام انه الدمع يذرف فرحاً.

وأتوجه بالشكر لكل من درسني أو ساعد وإلى مشرفتي الدكتورة (اميرة عبدالواحد فنجان)، وكل الأساتذة الذين يرجع لهم الفضل بعد الله سبحانه وتعالى في تلقيني (قسم الرياضيات)، كما اقدم الشكر والتقدير للأساتذة المشرفين على هذا البحث، والشكر موجه أيضا لإدارة جامعة (بابل) لحسن توفيرهم الخدمات للطلاب وتسهيلها ومساعدتهم في كل الأمور التي من شانها أن تمنحهم فضاءً مريحا للدراسة وطلب العلم في نظام وأمان .

المحتويات

Abstract

Clustering technology has important applications in data mining, pattern recognition, machine learning and other fields. However, with the explosive growth of data, traditional clustering algorithm is more and more difficult to meet the needs of big data analysis. How to improve the traditional clustering algorithm and ensure the quality and efficiency of clustering under the background of big data has become an important research topic of artificial intelligence and big data processing. The density-based clustering algorithm can cluster arbitrarily shaped data sets in the case of unknown data distribution. DBSCAN is a classical density-based clustering algorithm, which is widely used for data clustering analysis due to its simple and efficient characteristics. The purpose of this paper is to study DBSCAN clustering algorithm based on density. This paper first introduces the concept of DBSCAN algorithm, and then carries out performance tests on DBSCAN algorithm in three different data sets. By analyzing the experimental results, it can be concluded that DBSCAN algorithm has higher homogeneity and diversity when it performs personalized clustering on data sets of non-uniform density with broad values and gradually sparse forwards. When the DBSCAN algorithm's neighborhood distance eps is 1000, 26 classes are generated after clustering.

Chapter One

Introduction

1-1. Introduction

Clustering is the task of dividing the unlabeled data or data points into different clusters such that similar data points fall in the same cluster than those which differ from the others. In simple words, the aim of the clustering process is to segregate groups with similar traits and assign them into clusters. Clustering is a type of unsupervised learning method of machine learning. In the unsupervised learning method, the inferences are drawn from the data sets which do not contain labelled output variable. It is an exploratory data analysis technique that allows us to analyze the multivariate data sets. Clustering analysis is an unsupervised learning method that separates the data points into several specific bunches or groups, such that the data points in the same groups have similar properties and data points in different groups have different properties in some sense. Clustering analysis is broadly used in many applications such as market research, pattern recognition, data analysis, and image processing. Clustering can also help marketers discover distinct groups in their customer base. And they can characterize their customer groups based on the purchasing patterns.

1-2. Motivation

DBSCAN is a clustering method that is used in machine learning to separate clusters of high density from clusters of low density.[1]

1-3. Why does clustering work?

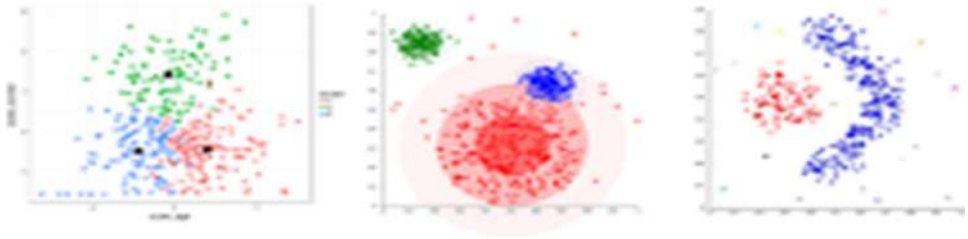


Figure 1-1: Clustering Work

Hierarchical clustering algorithm works by iteratively connecting closest data points to form clusters. Initially all data points are disconnected from each other; each data point is treated as its own cluster. Then, the two closest data points are connected, forming a cluster as figure (1-1) [2].

1-4. What is the cluster?

Cluster is the collection of data objects which are similar to one another within the same group (class or category) and are different from the objects in the other clusters. Clustering is an unsupervised learning technique in which there is predefined classes and prior information which defines how the data should be grouped or labeled into separate classes.

- Centroid-based Clustering.
- Density-based Clustering.
- Distribution-based Clustering.
- Hierarchical Clustering.[3]

1-5. Classification vs clustering

Classification is a supervised learning and the model learns a method for predicting the instance class from a pre-labeled (classified) instances while clustering is an unsupervised learning and the model tries to find "natural" grouping of instances for a given unlabeled data. Clustering allows us to find hidden relationship between the data points in the dataset.

Examples:

1. In marketing, customers are segmented according to similarities to carry out targeted marketing.

2. Given a collection of text, we need to organize them, according to the content similarities to create a topic hierarchy

3. Detecting distinct kinds of pattern in image data (Image processing). It's effective in biology research for identifying the underlying patterns.

How do we define good Clustering algorithms?

High quality clusters can be created by reducing the distance between the objects in the same cluster known as intra-cluster minimization and increasing the distance with the objects in the other cluster known as inter-cluster maximization.

Intra-cluster minimization: The closer the objects in a cluster, the more likely they belong to the same cluster.

Inter-cluster Maximization: This makes the separation between two clusters. The main goal is to maximize the distance between 2 clusters.

What is the best clustering algorithms

Top 10 clustering algorithms (in alphabetical order):

1. BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies)
2. DBSCAN (Density-Based Spatial Clustering of Applications with Noise)
3. Gaussian Mixture Models (GMM)
4. K-Means.
5. Mean Shift Clustering.
6. Mini-Batch K-Means.
7. OPTICS.
8. Spectral Clustering.[4]

Chapter Two

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and K-means Clustering

2-1.Density-Based Clustering Algorithms

Density-Based Clustering refers to unsupervised learning methods that identify distinctive groups/clusters in the data, based on the idea that a cluster in data space is a contiguous region of high point density, separated from other such clusters by contiguous regions of low point density.

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is a base algorithm for density-based clustering. It can discover clusters of different shapes and sizes from a large amount of data, which is containing noise and outliers.

The DBSCAN algorithm uses two parameters as figure (2-1):

- **minPts:** The minimum number of points (a threshold) clustered together for a region to be considered dense.
- **eps (ε):** A distance measure that will be used to locate the points in the neighborhood of any point.

These parameters can be understood if we explore two concepts called Density Reachability and Density Connectivity.

Reachability in terms of density establishes a point to be reachable from another if it lies within a particular distance (eps) from it.

Connectivity, on the other hand, involves a transitivity based chaining-approach to determine whether points are located in a particular cluster. For example, p and q points could be connected if p->r->s->t->q, where a->b means b is in the neighborhood of a.

There are three types of points after the DBSCAN clustering is complete:

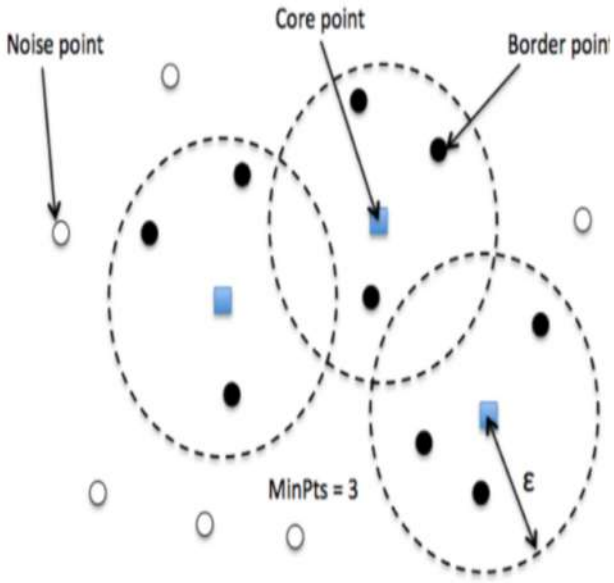
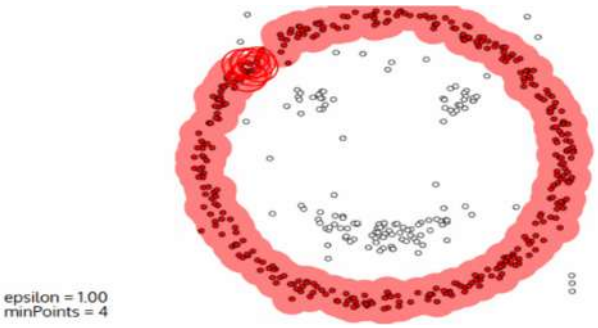


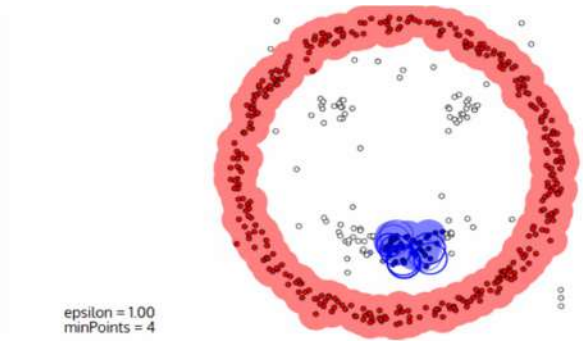
Figure 2-1: DBSCAN Technique

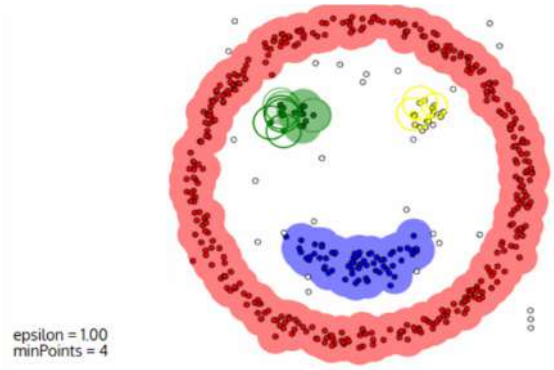
- **Core** — This is a point that has at least *m* points within distance *n* from itself.
- **Border** — This is a point that has at least one Core point at a distance *n*.
- **Noise** — This is a point that is neither a Core nor a Border. And it has less than *m* points within distance *n* from itself.[5]

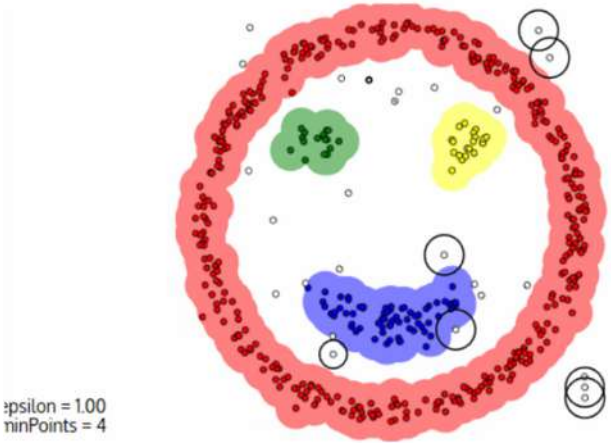
2-2. Algorithmic steps for DBSCAN clustering

- The algorithm proceeds by arbitrarily picking up a point in the dataset (until all points have been visited) as figure (2-2) .
- If there are at least 'minPoint' points within a radius of 'ε' to the point then we consider all these points to be part of the same cluster.
- The clusters are then expanded by recursively repeating the neighbourhood calculation for each neighbouring point [6][7].


epsilon = 1.00
minPoints = 4


epsilon = 1.00
minPoints = 4


epsilon = 1.00
minPoints = 4


epsilon = 1.00
minPoints = 4

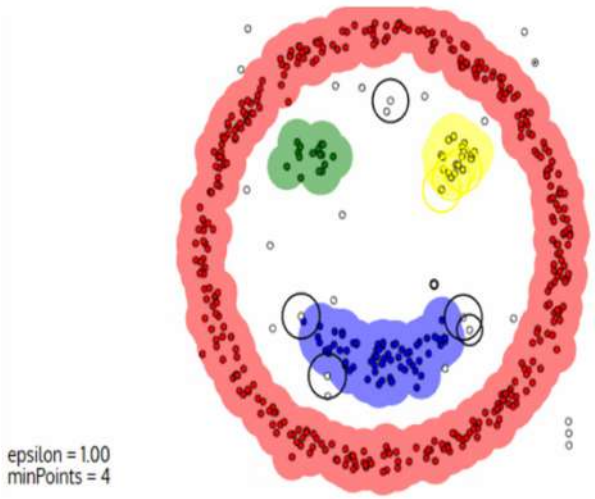


Figure 2-2: DBSCAN Clustering Steps

2-3. Why do we need a Density-Based clustering algorithm like DBSCAN when we already have K-means clustering?

K-Means clustering may cluster loosely related observations together. Every observation becomes a part of some cluster eventually, even if the observations are scattered far away in the vector space. Since clusters depend on the mean value of cluster elements, each data point plays a role in forming the clusters. A slight change in data points might affect the clustering outcome. This problem is greatly reduced in DBSCAN due to the way clusters are formed. This is usually not a big problem unless we come across some odd shape data. Another challenge with k-means is that you need to specify the number of clusters ("k") in order to use it. Much of the time, we won't know what a reasonable k value is a priori.

What's nice about DBSCAN is that you don't have to specify the number of clusters to use it. All you need is a function to calculate the distance between values and some guidance for what amount of distance is considered "close". DBSCAN also produces more reasonable results than k-means across a variety of different distributions. Below figure (2-3) illustrates the fact:[8],[9]

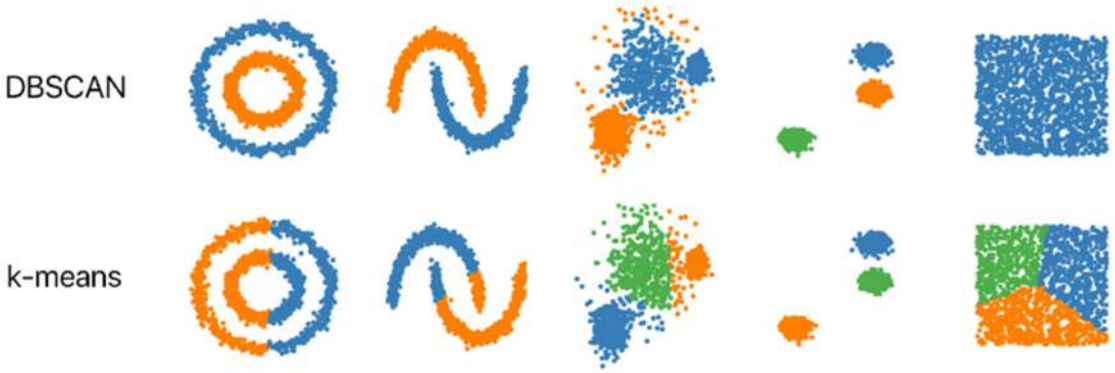


Figure 2-3: Comparison between DBSCAN and K-means Clustering

2-4. What is the difference between K-Means and DBSCAN?

- **K-Means**

K-means clustering is the partitioning algorithm. K-means recreates each data in the dataset to only one of the new clusters formed. A data or data point is assigned to the adjacent cluster using a measure of distance or similarity.

In k-means, an object is generated to the nearest center. It can define cannot-link constraints, and it modifies the center assignment process in k-means to the closest applicable center assignment.

When the objects are created to centers in sequence, at each step it can provide the assignments so far do not disorganize some cannot-link constraints. An object is created to the closest center therefore the assignment respects some cannot-link constraints [6][10].

- **DBSCAN**

DBSCAN represents Density-Based Spatial Clustering of Applications with Noise. It is a density-based clustering algorithm. The algorithm improves regions with adequately

high density into clusters and discovers clusters of arbitrary structure in spatial databases with noise. It defines a cluster as a maximum set of density-connected points.

A density-based cluster is a set of density-connected objects that is maximal regarding density-reachability. Each object not contained in some cluster is considered to be noise.

DBSCAN checks for clusters by checking the ε-neighborhood of every point in the database. If the ε-neighborhood of a point p contains more than MinPts, a new cluster with p as a core element is produced. DBSCAN iteratively assemble precisely density-reachable objects from these essential element, which can include the merge of a few density-reachable clusters. The process eliminates when no new point can be added to any cluster Table (2-1) shows that [12] [13] .

Table 2-1. Comparison between K-Means and DBSCAN.

K-Means	DBSCAN
K-means generally clusters all the objects.	DBSCAN discards objects that it defines as noise.
K-means needs a prototype-based concept of a cluster.	DBSCAN needs a density-based concept.
K-means has difficulty with non-globular clusters and clusters of multiple sizes.	DBSCAN is used to handle clusters of multiple sizes and structures and is not powerfully influenced by noise or outliers.
K-means can be used for data that has a clear centroid, including a mean or median.	DBSCAN needed that its definition of density, which depends on the traditional

K-Means	DBSCAN
	Euclidean concept of density, be significant for the data.
K-means can be used to sparse, high dimensional data, including file data.	DBSCAN generally implements poorly for such information because the traditional Euclidean definition of density does not operate well for high dimensional data.
The basic K-means algorithm is similar to a statistical clustering approach (mixture models) that consider all clusters come from spherical Gaussian distributions with several means but the equal covariance matrix.	DIISCAN creates no assumption about the distribution of the record.[10]

2-5. Pros and Cons of DBSCAN

Pros:

- Does not require to specify number of clusters beforehand.
- Performs well with arbitrary shapes clusters.
- DBSCAN is robust to outliers and able to detect the outliers.

Cons:

- In some cases, determining an appropriate distance of neighborhood (eps) is not easy and it requires domain knowledge.

- If clusters are very different in terms of in-cluster densities, DBSCAN is not well suited to define clusters. The characteristics of clusters are defined by the combination of eps-minPts parameters. Since we pass in one eps-minPts combination to the algorithm, it cannot generalize well to clusters with much different densities [14].

2-6. Conclusion

Density-based clustering algorithms can learn clusters of arbitrary shape, and with the level set tree algorithm, one can learn clusters in datasets that exhibit wide differences in density.

However, I should point out that these algorithms are somewhat more arduous to tune contrasted to parametric clustering algorithms like K-Means. Parameters like the epsilon for DBSCAN or for the level set tree are less intuitive to reason about compared to the number of clusters parameter for K-Means, so it’s more difficult to choose good initial parameter values for these algorithms.

References

References:

1. Ester, Martin; Kriegel, Hans-Peter; Sander, Jörg; Xu, Xiaowei (1996). Simoudis, Evangelos; Han, Jiawei; Fayyad, Usama M. (eds.). A density-based algorithm for discovering clusters in large spatial databases with noise (PDF). Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96). AAAI Press. pp. 226–231. CiteSeerX 10.1.1.121.9220. ISBN 1-57735-004-9.
2. ^ "Microsoft Academic Search: Papers". Archived from the original on April 21, 2010. Retrieved 2010-04-18. Most cited data mining articles according to Microsoft academic search; DBSCAN is on rank 24.
3. ^ "2014 SIGKDD Test of Time Award". ACM SIGKDD. 2014-08-18. Retrieved 2016-07-27.
4. ^ Jump up to:[a b c d e f g h i j k l] Schubert, Erich; Sander, Jörg; Ester, Martin; Kriegel, Hans Peter; Xu, Xiaowei (July 2017). "DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN". ACM Trans. Database Syst. 42 (3): 19:1–19:21. doi:10.1145/3068335. ISSN 0362-5915. S2CID 5156876.
5. ^ "TODS Home". tods.acm.org. Association for Computing Machinery. Retrieved 2020-07-16.
6. ^ Jump up to:[a b] Ling, R. F. (1972-01-01). "On the theory and construction of k-clusters". The Computer Journal. 15 (4): 326–332. doi:10.1093/comjnl/15.4.326. ISSN 0010-4620.
7. ^ Jump up to:[a b c d] Sander, Jörg; Ester, Martin; Kriegel, Hans-Peter; Xu, Xiaowei (1998). "Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and Its Applications". Data Mining and Knowledge Discovery. Berlin: Springer-Verlag. 2 (2): 169–194. doi:10.1023/A:1009745219419. S2CID 445002.

8. ^ Jump up to:[a] [b] [c] Campello, Ricardo J. G. B.; Moulavi, Davoud; Zimek, Arthur; Sander, Jörg (2015). "Hierarchical Density Estimates for Data Clustering, Visualization, and Outlier Detection". ACM Transactions on Knowledge Discovery from Data. 10 (1): 1–51. doi:10.1145/2733381. ISSN 1556-4681. S2CID 2887636.
9. ^ Kriegel, Hans-Peter; Kröger, Peer; Sander, Jörg; Zimek, Arthur (2011). "Density-based Clustering". WIREs Data Mining and Knowledge Discovery. 1 (3): 231–240. doi:10.1002/widm.30. S2CID 36920706. Archived from the original on 2016-11-17. Retrieved 2011-12-12.
10. ^ Schubert, Erich; Hess, Sibylle; Morik, Katharina (2018). The Relationship of DBSCAN to Matrix Factorization and Spectral Clustering (PDF). Lernen, Wissen, Daten, Analysen (LWDA). pp. 330–334 – via CEUR-WS.org.
11. ^ Sander, Jörg (1998). Generalized Density-Based Clustering for Spatial Data Mining. München: Herbert Utz Verlag. ISBN 3-89675-469-6.
12. ^ Campello, R. J. G. B.; Moulavi, D.; Zimek, A.; Sander, J. (2013). "A framework for semi-supervised and unsupervised optimal extraction of clusters from hierarchies". Data Mining and Knowledge Discovery. 27 (3): 344. doi:10.1007/s10618-013-0311-4. S2CID 8144686.
13. ^ Kriegel, Hans-Peter; Schubert, Erich; Zimek, Arthur (2016). "The (black) art of runtime evaluation: Are we comparing algorithms or implementations?". Knowledge and Information Systems. 52 (2): 341. doi:10.1007/s10115-016-1004-2. ISSN 0219-1377. S2CID 40772241.
14. Density-based clustering algorithms – DBSCAN and SNN by Adriano Moreira, Maribel Y. Santos and Sofia Carneiro

وزارة التعليم العالي والبحث العلمي
جامعة بابل
كلية التربية للعلوم الصرفة
قسم الرياضيات

التجميع المكاني القائم على الكثافة للتطبيقات ذات الضوضاء (DBSCAN)

بحث مقدم لمجلس جامعة بابل ، كلية العلوم الصرفة ، قسم الرياضيات ،
من متطلبات الحصول على درجة البكالوريوس

الطالبة
زهراء كاظم خليف

الاستاذة
أميرة عبد الواحد فنجان

١٤٤٤ هـ ٢٠٢٣ م