Turkish Journal of Computer and Mathematics Education Vol.12 No.13 (2021), 5391-5402

Research Article

(Density Based Spatial Clustering for Noisy Gene Expression Data)

Awrad Dawood Salman & Basad Al-sarray

Awrad.smart@gmail.com
Computer Science Department, Collage of Science, University of Baghdad, Baghdad, Iraq

Abstract

The Data Mining is about information examination methods. It is helpful for extricating covered up and fascinating examples from huge datasets, when it comes to extracting information from a large volume of spatial data collected from a variety of sources, grouping methods are critical. A pioneering thickness-based approximation is Density Based Spatial Clustering of Applications with Noise. It can locate groups of any arbitrary size and shape in data bases that include even commotion and exceptions. This study shows a detailed definition of DBSCAN works through different sheets of the most popular pattern presented up until now.

Keywords: Data Mining, Clustering, DBSCAN, Swarm, cuckoo, gene expiration.

1. Introduction

Data mining has recently become popular as a means of extracting valuable patterns and information from data. Using unsupervised techniques like clustering and supervised techniques like classification, these approaches discover and detect valuable information from results. Clustering techniques indicated collecting data or objects due to common standards, which is accomplished by a collection of related items simply a set of standards [1]. Clustering algorithms divide objects into numbers of smaller clusters to produce various Clusters of group stakeholders, but almost similar members of each cluster. Clustering techniques are thus data mining methods that are used in a range of applications such as flow shop scheduling, image processing, wireless sensor networks, intrusion detection, agriculture data, industrial data analysis, financial classifications, customer relationship management, bioinformatics, and dematrix, in which various approaches have been utilized. There are a number of clustering algorithms successfully used in real-life data mining problems and fulfill the requirements [2]. Nevertheless, lots of limitations are present in the majority of the current clustering techniques. In general, there are four types of clustering algorithms: partitioning methods, hierarchical methods, density-based methods, and grid-based methods. Partitioning algorithms divide datasets into distinct clusters without overlapping members. The K-means algorithm, which clusters large datasets based on the shortest runtime, is the most commonly used algorithm in the field.

2. Clustering:

Clustering algorithms can be divided into four categories: partitioning methods, hierarchical methods, density-based methods, and grid-based methods. Partitioning algorithms separate datasets into distinct clusters with no members in common. The most widely used algorithm in the field is the K-means algorithm, which clusters large datasets based on the shortest runtime. [17].Clustering is an unfettered learning method that associates data, objects or patterns based on measures of similarity .In the Rd space, and we might come across objects like data points. Patterns belonging to a particular class have greater similarity between them than similar patterns belonging to a different class [1-3]. Block analysis is performed to improve understanding of data in context, eg, compiling relevant documents for browsing, finding protein and gene structures with similar functions, or as a method of data compression [4]. A large number of pattern assembly techniques have been developed Analysis, document retrieval, clustering, decision-making, image segmentation, data mining, and so far there remains a major challenge in defining clusters precisely. Clustering approaches are broadly categorized into partial, hierarchical, and density-based methods (see Figure 1). [1].

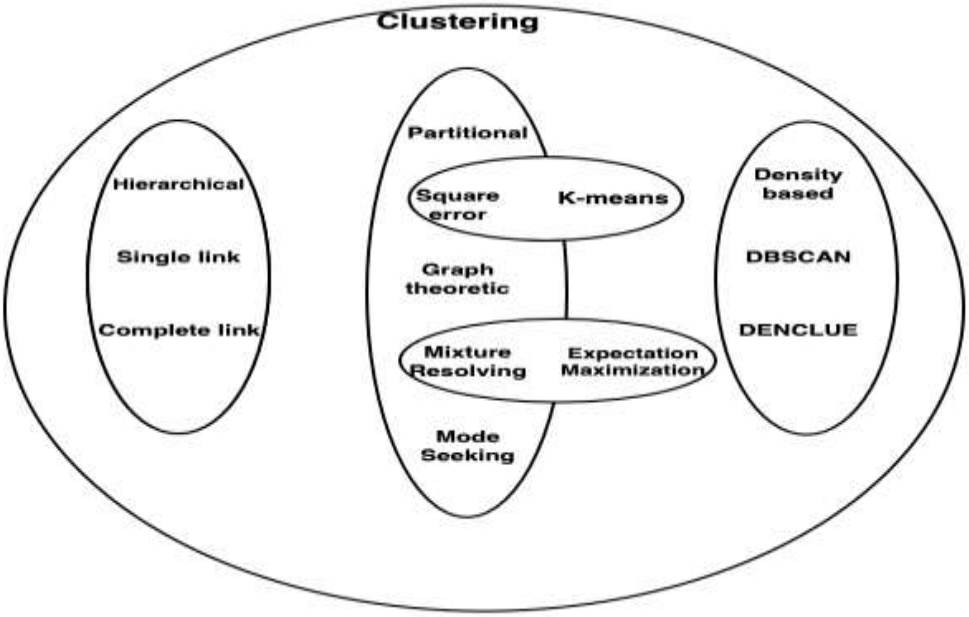

Figure 1 Clustering types and approaches

3. Clustering Algorithms:

Clustering algorithms are commonly used in mining of data systems to find similarities in the underlying data. The majority of conventional clustering algorithms are only capable of managing datasets with continuous or categorical attributes. In real-world data mining issues, Mixed-attribute datasets, on the other hand, are popular. BIRCH is a clustering approach that is particularly well suited to very broad datasets. Partial clustering approach contains squared error method, eg, K-mean algorithm, graph theory assembly, separation of components, search for a method [1].Hierarchical clustering gives us a branched diagram An overlapping set of patterns, an example is the chameleon [5]. The hierarchical method adopts a group or partitioning policy to define the cluster. The concept behind density-based clustering is to determine the area's density. The purpose of DBCLAs is to find diverse levels of clustered aggregate granules with appropriate noise identification. The idea of density used by DBCLAs enables the combined regions of the data space to be separated from the noise. In DBCLAs, clusters are expressed as spaces with a higher density than the rest of the data space [6]. DBCLAs make it easy to discover groups of arbitrary shapes. Over the past two decades, a lot of density-based group techniques have been proposed. The aim of these methods is to obtain groups of densities that are relatively uniform across the data space. Other notable clustering patterns are: proximity-based clustering, fuzzy clustering, artificial neural network (ANN) -based clustering and kernel-based technologies [1]. The idea of evolutionary methods of aggregation, which uses a multitude of solutions to obtain the clearest and best disaggregation of data globally. Clustering algorithms come with their own set of challenges. Depending on the characteristics of the data and the mechanism adopted to form clusters.

4. The clustering of applications with noise using density-based spatial clustering:

This type of problem is solved using the density based binding method (DBSCAN) [32-36]. Low-density points are separated by regions using groups that are described and allocated to dense areas in the data space. The DBSCAN algorithm is based on the concept of "clusters and noise." Essentially, any point on the block must encompass a neighborhood of at least a certain radius. A certain number of points is needed. The DBSCAN algorithm enforces a collection of points in a dedicated space while also grouping related data points in a nearby area (points with many close neighbors). They are viewed as external (noise) points that are found in low-density areas on their own. Aside from spatial data sets, Massive quantities of data with smaller cluster patterns relating to various dimensions. This

process consumes less calculation time than others. Shapes of image data sets should be adjusted for poor cases that are always repetitive and must be extracted. DBSCAN implements a data clustering algorithm that regards the process of density-based clustering through the evaluation of data location and distribution. A group of points in a space is grouped together with the concept of density. During the process of applying the algorithm, she grouped closely close sample points. Those that were alone in low-density areas (whose closest neighbors were very far away) were also marked as outdoor points. DBSCAN is a common clustering algorithm for misconfigured sample data, particularly in earth sciences and image classification. DBSCAN is unfamiliar to the majority of our readers. DBSCAN can sort data into groups of various formats as well, which is another powerful feature. DBSCAN operates in the following way:

• DBSCAN divides a dataset into n dimensions by drawing an n-dimensional figure around each data point in the dataset and counting the number of data points. DBSCAN considers this pattern to be a block. DBSCAN usually works by enlarging the block, going through each and every point in the cluster, and counting how many other data points are nearby.

The DBSCAN algorithm is illustrated in Figure 2 with a simple implementation.

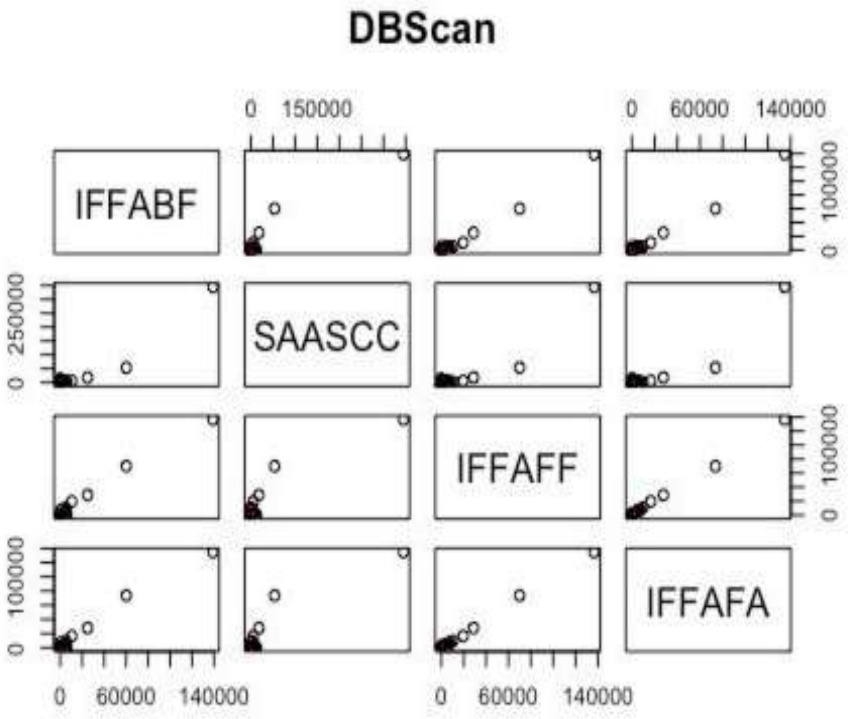

Figure 2 simple implementation of DBSCAN algorithm

• DBSCAN Benefits and Functionality

• Within a data set, it can be used to separate high-density and low-density groups. In a dataset, knows how to handle outliers.

DBSCAN Gaps and Disadvantages

• While DBSCAN shines at distinguishing high-density clusters from low-density clusters, it also spends a lot of time looking for solutions that have clusters of similar density.

• Problems with data with a lot of dimensions. DBSCAN excels at transforming data into different dimensions and shapes. DBSCAN, on the other hand, can only go so far; if you send it too many data with too many dimensions, it will fall behind in the study.

5. Density-based spatial clustering of applications with noise (DBSCAN) Methods and Approaches :

Many studies and research have been done on (Density dependent Spatial clustering for noisy gene expression data), including:

• **Guang Feng et.al 2020[7].** They developed a method that combines particle swarm optimization, non-dominant sorting, and multi-classifier techniques including the k-nearest neighbor method, quick decision tree, and kernel density estimation." Bayes' theorem is used to revise the results in order to arrive at the most accurate breast cancer prediction. The proposed particle swarm optimization and non-domination sorting with classifier technique model will assist in selecting the most important breast cancer prediction functions. The features chosen decide the problem model's goal. This model was tested using the "WBCD and WDBC" breast cancer data sets from the UCI machine learning data repository. Precision, accuracy, time complexity, and sensitivity are all taken into consideration.

• **Nawel Zemmal et.al 2020[5].** A hybrid paradigm incorporating "active learning (AL)" and "particle swarm optimization (PSO)" algorithms is proposed to reduce the cost of labeling while increasing the effectiveness of the classifier. Eighteen (18) benchmark datasets were used to compare the proposed solution to three well-known classifiers from different learning paradigms: "AL–NB an active learning algorithm" using "Nave Base classifier" and "Margin Sampling strategy," "SVM (Support Vector Machine)", "ELM (Extreme Learning Machine)" with supervised learning, and "TSVM (Transductive Support Vector Machine)" with unsupervised learning. Experiments have shown that the proposed method would save clinicians time and effort while annotating medical data to create a reliable classifier. To speed up the time-consuming process of tagging, researchers used active learning. PSO, a nature-inspired algorithm, employs a novel uncertainty measure to choose the most informative medical instances from a large number of unlabeled instances while also improving the classifier's accuracy.

• **Uzma et.al 2020[8].** Principal component analysis, correlation, and spectral-based feature selection are used to implement three filtering techniques in the first stage of the classification cancer specimen. Using auto encoder-based clustering, the genetic algorithm is used to analyze the chromosome. The classification task can then be applied to the function subset that has been developed. Help vector machine, k-nearest neighbors, and random forest were used to solve the problem of single-classifier dependence. Six benchmark gene expression datasets were used to evaluate the results. The comparison was then made using four state-of-the-art equivalent algorithms. Three sets of experiments are conducted to evaluate these tests: evaluation of selected features based on sample-based clustering, modification of optimal parameters, and selection of a better performing classifier. The distinction is made with the help of accuracy, memory, false positive rate, precision, F-measure, and entropy.

• **Chun Guan et.al 2019[4].** The DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm locates arbitrary formed clusters in a dataset. DBSCAN has three flaws: first, the parameters are difficult to set; second, users cannot control the number of clusters; and third, DBSCAN cannot be used as a classifier directly. To fix the shortcomings of DBSCAN, this paper proposes a new "particle swarm Optimized Density-based Clustering

and Classification (PODCC)" process. PSO is a well-known Evolutionary and Swarm Algorithm (ESA) that has been used to solve a number of optimization problems, including data analytics. Users set the number of input clusters to PODCC using the proposed fitness function. Twenty datasets (10 synthetic and 10 benchmarking) from various open sources were used to validate the proposed process.

- **Asgarali Bouyer& Abdolreza Hatamlou 2018[1]**. The K-means algorithm divides objects into smaller, disjoint classes with the most resemblance to objects in the same category and the most dissimilarity to objects in other categories in partitioned data clustering. KHarmonic Means (KHM) was used in conjunction with an improved Cuckoo Search (ICS) and PSO to create a clustering algorithm. The aim of ICS is to use the Levy flight method to find the global optimum approach for dynamically and intelligently changing the radius. As a consequence, it is quicker than a standard cuckoo hunt. PSO affects ICS so that it does not collapse into local optima. The proposed algorithm, dubbed ICMPKHM, solves the KHM local optima problem with significantly improved effectiveness and stability.

- **Santosh Kumar Majhi&Shubhra Biswal 2018 [2].** A hybrid clustering approach using K-means and Ant Lion Optimization (ALO) has been proposed for optimal cluster analysis. ALO stands for "adversarial learning optimization." The performance of the "KMeans-PSO, Revised DBSCAN, DBSCAN, KMeans-FA, and K-Means" clustering approaches is compared to the output of the suggested algorithm based on various efficiency parameter metrics. Statistical analysis was performed on eight datasets. In terms of F-measure and sum of intra cluster distances, the results showed that the hybrid of K-Means and ALO approach outperformed the other three algorithms.

- **Chun Guan 2018[9]**. Density-based clustering is a form of clustering that allows for the creation of arbitrary-shaped clusters. DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a well-known density-based clustering algorithm. Particle Swarm Optimization (PSO), Genetic Algorithms (GAs), Artificial Bee Colony (ABC), and Differential Evaluation are just a few of the ESAs that have been discussed so far (DE). In the ESA-DCC system, ESA is used to find the best density-based classification and clustering parameters to overcome DBSCAN's challenge. To compare the ESA-DCC methods to DBSCAN and K-means, ten datasets will be used.

- **Limin Wang et.al 2018[10].** It uses the cuckoo search algorithm to implement a parameter adaptive density based spatial clustering of noisy applications, which provides a fast solution to the global optimization problem. The enhanced algorithm completes the clustering process automatically without the involvement of humans, thanks to the cuckoo search algorithm for determining the best global parameter Eps. The simulation results show that this algorithm is capable of selecting a suitable Eps parameter value and accurately reaching clustering process results.

Table 1: summary of current approaches and dataset

Title	Techniques	Conclusion	data sets
Multi-modal prediction of breast cancer using non-dominating sorting with PSO	PSO using non-dominating method Bayes' theorem	98.28% and 98.8% accuracy were achieved by PSO-NDS for features between 5–6 and 15–20	WBCD and WDBC data sets
Using Projection-Based Clustering to generate Distance and Density-Based Clusters in High-Dimensional Data	PBC projection methods k-means	There are six parts in total. The first three are concerned with cluster propensity (also known as clusterability).	Comprised of 12 databases containing well-known classifications. (Thrun and Ultsch 2020a) available in the R package "FCPS" on CRAN

3. Gene encoder: a feature selection technique through unsupervised deep learning-based clustering for large gene expression data	deep/machine learning methods k-Means clustering algorithm gene expression Random Forest (RF) Fitness function	the framework: (a) assess the chosen features through the use of clustering based on samples, (b) analyzing the parameter setting, i.e., TopN and mutation rate, and (c) evaluating three classifiers, i.e., K-NN, RF, and SVM. Table 2 shows the parameter settings used all experiments. DLBCL type cancer.	DLBCL, Lung and colon cancers, leukemia, and center nerves system.
4. Particle Swarm Optimization Based Swarm Intelligence for Active Learning Improvement: Application on Medical Data Classification	Particle swarm optimization Binary particle swarm optimization (BPSO)	Details of the parameters that take part constructing the AL-PSO scheme. The process of AL-PSO starts with a starting set L0 including 10% only from the labeled data selected at random from the datasets "Indicates the classification performances (mean of the 10-fold cross validation and the standard deviation) of the proposed AL–PSO. The current results are compared with those of various classifiers based on distinct learning paradigms through the use of 18 datasets for validation.	Echocardiogram (ECG), catheterization diagnostic, Breast Cancer (Ljubljana), Diabetic Retinopathy Debrecen (DRD), Cleveland heart disease, Breast Cancer Wisconsin (WDBC), heart disease, cardiac, Pima Indians Diabetes, Parkinson's, and hepatitis are among the nine medical datasets selected from the UCI Machine Learning Repository (ECG).
An Efficient Hybrid Clustering Method based on "Improved Cuckoo Optimization" and	K-harmonic means algorithm Cuckoo Search via Lévy flight	On each dataset, each algorithm was run 100 times. Table 3 shows the simulation results for total of squared errors (best, average, and worst solutions	ArtSet1 (n = 300, d = 2, k = 3): an artificial dataset was included with a two-featured problem and three unique classes.

Modified "PSO" Algorithms	Standard PSO and Modified PSO algorithms hybrid data-clustering algorithm "Cuckoo Search algorithm "ICMPKHM Clustering Algorithm"	among 100 runs), solution standard deviation, and F-measure. KHM-IPSO and PSOKHM are less accurate than ICMPKHM. In addition, with the exception of the cancer data collection, the suggested algorithm has the lowest standard deviation of all the other algorithms (for PSOKHM and KHM-IPSO). The Iris dataset, for example, revealed that ICMPKHM converges to the global optimum.	Iris Data set (N = 150, d = 4, K = 3): it It's worth noting that this is perhaps the most well-known database." Data set for wine (N = 178, d = 13, K = 3): This information comes from the Institute of Pharmaceutical and Food Research and Technologies in Italy, which conducted a chemical analysis of wines grown there.
PSO for feature selection and bandwidth determination of kernel density estimation based classifiers in diagnosis of breast cancer	genetic algorithms ant colony optimization (ACO) PSO artificial neural networks (ANNs) PSO-KDE	The PSO-KDE model with those of GA-KDE. Afterwards, a comparison was held between the results of PSO-KDE and those of other classifiers addressed in the literature Once the feature subset and optimal kernel bandwidth were established using PSO, the experiments were carried out on the testing set using the selected feature subset and determined kernel bandwidth.	WBCD WDBC

6. Gene expression:

Gene expression is a tightly controlled process that regulates the structure and both living cells have the potential to adapt. It involves the conversion of gene information into a functioning gene product, which may be a protein or a non-coding gene such as small nuclear RNA (snRNA) or transfer ribonucleic acid (tRNA). Since this mechanism is used by all known life, the regulation of gene expression is the most important aspect of learning the genotype: phenotype interaction. Transcription, RNA splicing, translation, and posttranslational modification are the four major stages in the development of proteins. During transcription, the "RNA" polymerase complements copies of a single strand mRNA from one strand of a template "DNA." During transcription, introns are deleted from the sequence, resulting in certain changes to the "mRNA," i.e., "RNA" splicing. The new "mRNA" is then used as a blueprint for putting together a protein chain. Gene expression researches the sum of transcribed "mRNA" in a biological system and how it affects cell function. [24].

7. Measure of proximity to gene expression data

The similarity (or distance) between each two data objects is measured by a rough metric. Gene expression data objects, regardless of genes or samples, can be formulated as numerical Vector ~ Oi. foijj1 _ j _ pg, where oij represents the feature value of jth of the data object i and p is the number of features. The proximity between two objects Oi and Oj is The proximity function measured from the corresponding vectors ~ Oi and ~ Oj. One of the most common methods for determining the distance between two data objects is to use Euclidean distance. In dimension p, the distance between the points Oi and Oj. The term "space" is described as: Euclidean .00000(Oi,Oj) =√∑(Oid-ojd)2 Or features) [71]. This problem is handled through standardizing each vector with a "zero" mean and "one: variance before calculating the distance [66], [59], [56]. An alternative measure is the "Pearson Correlation Coefficient", which measures the similarity between two shapes Patterns of expression (metafiles). The Pearson correlation coefficient is calculated by auditing two data items, Oi and Oj. It is the medium for ~ Oi and ~ Oj, respectively. "Pearson's correlation coefficient" explains each object as a arbitrary variable with observations of p and measures the similarity between two objects by computing linearity .The relationship between the distributions of the corresponding random variables, The "Pearson Correlation Coefficient" is widely known and has proven effective as a measure of gene similarity in numerous experiments. Expression data [21], [24], [16]. However, the pilot study has shown that they are not accurate with respect to outliers [30], and thus are likely to yield false positives that specify a high degree of similarity for a pair of disparate patterns. If there are two formulas that have a common summit or valley at one feature, this feature will dominate the link, although the patterns in the remaining features may be very different. This observation sparked A modified scale on it is called the Jackknife Correlation [19], [30],Pearson correlation coefficient for coordinates Oi and Oj data with lth omitted feature. To avoid the "dominance effect" of individual outliers the Jackknife correlation has been used. More general versions can also be derived from the Jackknife link that are robust to more than one external. However, the generalized Jackknife link, which may involve enumerating various combinations of features to be deleted, will be computationally expensive and rarely used. They are robust for non-Gaussian distributions [14], [16]. The Spearman rank-order correlation coefficient was discovered as a measure of similarity to resolve this. Substitution yields the order connection. The level of a numerical expression with its order gets rid of all conditions. For example, get rid of. 3 If oid is the third-highest value among oik, where 1 _k _ p. The Spearman Link, The coefficient does not require the assumption of a Gaussian distribution and is robust and more correct against random values than the "Pearson Correlation Coefficient". Nevertheless, as a result of the arrangement, there will be a huge loose of information in data. Generally, our experimental results indicate that the "Pearson rank-order correlation coefficient" performs better than "Spearman rank-order correlation coefficient"

.**Table 2:** The table represents comparisons based on data set according to previous research and whether it contains the gene expression or not, as well as the type and class of data set used.

Name of paper	No. of Data set	type	Use Gene expiration
Multi-modal breast cancer prediction that uses PSO with non-dominating sorting, 2020.	569 WDBC	images	
Using Projection-Based Clustering to Find Distance and Density-Based Clusters in High-Dimensional Data2020.	699 cases	images	
Gene encoder: a technique to select features through "unsupervised deep learning-based clustering" for significant gene expression data2020.	The leukemia datasets 3571	Numeric	
"PSO Based Swarm Intelligence for Active Learning Improvement: Application on Medical Data Classification.2020.	18 bench-mark datasets with balanced data	Numeric	X
Fault diagnosis of rolling bearing that uses symmetrized dot pattern and density-based clustering.2019	600	Signals, images	X

A very good Hybrid Clustering Method based on developed PSO and Cuckoo Optimization Modified Algorithms.2018	300 patterns	Iris	X
IoMT-based computational approach for detecting brain tumor.2020	62 MR	images	X
Identification of cell types from single cell data using stable clustering.2020	eight publicly available scRNA-seq datasets	single cell	
Particle swarm optimization for bandwidth determination and feature selection of kernel density estimation based classifiers in diagnosis of breast cancer.2015	699	images	
Adaptive attenuation correction during H-scan ultrasound imaging using K-means cluste.2019	256	Images of sound signal	X
An enhanced density-based spatial clustering of application that include noise. 2018		UCI Cluster data-set	

8. Difficulties and challenges

The density-based spatial clustering of applications with noise (DBSCAN) algorithm has always had the following issues: first, experience completely controls the threshold determination (minPts, Eps), lowering the quality of the clustering results; second, this algorithm is ineffective at processing large amounts of data. The method of setting the threshold parameters in the classical algorithm is specifically set to minPts 14 4, and the Eps value was calculated by the observation, The accuracy of the follow-up clustering results cannot be assured because of this. In recent years, a significant number of improvements have been made to solve problems with clustering efficiency [3]. To fix the problem of the "DBSCAN" algorithm output being dependent on two defined parameters, epsilon and minPts, a new concept in core density reachable based on the influence space was proposed. This algorithm reduces the harmonic average of all evolutionary algorithms when combined with conventional clustering algorithms, which is an alternative to limiting their flows. Swarm intelligence and meta-heuristic algorithms like PSO, DB scan algorithm, Genetic Algorithm, and regular Cuckoo Search have been used in recent years. DBSCAN, on the other hand, is considered to have a number of flaws, including I the need for client participation in determining boundary esteems before executing the calculation; (ii) the difficulty in selecting substantial bunches from datasets with varying densities; and (iii) the computational complexity. To overcome these drawbacks, several experts attempted to improve the basic DBSCAN calculationTo solve and reduce noise and data loss, we use more than algorithms like Swarm, Cuckoo, and DBScan. To solve the problems above, a global optimization algorithm that can find the correct answer in a reasonable amount of time is needed. Met heuristics have risen in popularity in recent years as an efficient and suitable tool for solving global optimization problems. Stochastic processes are the most well-known, in which the initial swarm is equated to all of its neighbors while retaining the best output each time. PSO, a subfield of computational intelligence, is related to swarm intelligence and collective intelligence. The PSO algorithm is a tool of adaptive analysis. The search starts with a swarm of particles, or solutions. Each swarm particle represents a possible solution to the problem of optimization [4]. The particle is piloted in a multidimensional search space, and its location in that space shifts based on its and its neighbors' experiences. PSO intelligence is a low-cost programming tool that can be used for a number of tasks. The following parameters were used as measurement parameters: prediction rate, precision, sensitivity, specificity, and time complexity. [5]. the use of a multi-classifier reduces the possibility of errors in the output. PSO has been successfully used in a number of science, engineering, and numerical optimizations because it is easy to use and requires only a few parameters to be changed, and it provides good and consistent performance, has a robust search capability, does not require comprehensive knowledge of the given problem, and can be optimized for

nonconvex and multimodal problems. [6].

9. Most common Algorithms

9.1 The particle swarm algorithm (PSO):

The particle swarm algorithm begins with the development of initial particles and the assignment of initial velocities to them. It determines the most optimum location and best (lowest) function value by calculating the objective function at each particle position. It calculates new velocities based on the current velocity, the particle's best individual locations, and the particle's best neighboring locations. Then it iteratively updates particle positions, velocities, and neighbors (the current position is the previous position plus the velocity, modified to keep particles within bounds).Iterations continue until the algorithm meets a criterion for halting.

9.2 The Cuckoo algorithm:

This meta-heuristic optimization algorithm was recently created to solve optimization problems. It's based on brood parasitism in some cuckoo species, as well as Levy flights random walks, and it's inspired by nature. The cuckoo quest's parameters are typically kept constant for a fixed period of time, lowering the algorithm's efficiency. To address this problem, a proper method for fine-tuning the cuckoo search parameters must be established. Cuckoos are fascinating birds, not just because of their beautiful songs, but also because of their ruthless reproduction strategy. Some cuckoo birds, such as the Ani and Guira, lay their eggs in host bird nests and may remove other eggs in order to increase the chances of their own hatching. This article addressed cuckoo behavior and how they lay eggs in the nests of other host birds.

Table 3: According to previous research, the table represents comparisons based on the techniques and algorithms used within them, and the weaknesses points.

Name of paper	Techniques and algorithms used	Weaknesses
Using Projection-Based Clustering to Find Distance-and Density-Based Clusters in High-Dimensional Data	k-means, clustering algorithms (PBC)	Difficulty calculating experiment and projection variance of results
Gene encoder: a feature selection strategy used for large gene expression data through the use of unsupervised deep learning-based clustering.	Gene expression , Clustering , Genetic algorithm, component analysis (PCA)	The cluster of the set of informative features minimizes the objective function.
Particle Swarm Optimization Based Swarm Intelligence for Active Learning Improvement: Application on Medical Data Classification	Particle swarm optimization, global optimization problem, and binary particle swarm optimization are all terms used to describe particle swarm optimization (BPSO)	The most difficult problem in AL is determining whether or not a given instance is informative. The most popular method for selecting data from a large pool of unlabeled data is the uncertainty sampling method. That are to be classified by the expert, is one of the strategies that can solve this challenge.
Fault diagnosis of rolling bearing using symmetrized dot pattern and density-based clustering	Rolling bearing, Fault diagnosis, Symmetrized dot, Density-based clustering, (ASDP-DBSCAN)	Businesses take time, so developing a system to minimize calculation time and design more feature extraction

		and diagnostic methods is very exciting.
A very effective Hybrid Clustering Method based on Modified Particle Swarm Optimization and Improved Cuckoo Optimization Algorithms.	"Data Clustering", PSO, "K-Harmonic Means", "Cuckoo Optimization" Algorithm, "ICMPKHM" Clustering Algorithm.	One obstacle of ICMPKHM is its runtime, for achieving faster convergence, accuracy and runtime.
Multi-modal prediction of breast cancer utilizing "PSO" with non-dominating sorting	Swarm intelligence, feature selection, multi-classification, Bayes' theorem, particle swarm optimization	Subjecting datasets of significant amount of data to prediction and scanning decreases the accuracy and prediction-rate naturally.
"PSO" for bandwidth determination and feature selection of "kernel density estimation" based classifiers in diagnosis of breast cancer	Particle Swarm Optimization, PSO-KDE, GA-KDE model	A weakness in establishing a multipurpose method based on PSO for feature selection and bandwidth limitation in the kernel The classifier is based on density estimation which simultaneously reduces classification error and The number of features.
An improved density-based spatial clustering of application with noise	Density-based spatial clustering of applications with noise, Cuckoo search algorithm, CS-DBSCAN clustering algorithm.	this method lacks a to find a more rational and effective way to determine the MinPts Parameter.

10. Conclusion:

In this review paper, a number of recent works are presented in which the development of the DBScan algorithm has been explored. Where the researchers studied the advantages and disadvantages of the different techniques used in developing and integrating several techniques and algorithms to obtain better results and good features with less noise, and provided details of the methods used with error rates. In particular, the steps that Used to get the best results and in less time The previous research included several techniques, It includes the use of the DBScan algorithm with the Swarm algorithm, as well as the cuckoo with the use of several types of data set, and classification if the data set including the gene expression or not, and using a huge and complex data that are processed and applied to the algorithm to obtain the best results through the addition and development the DBScan algorithm to get good feature with less time.

References

1. Asgarali Bouyer& Abdolreza Hatamlou 2018. An Efficient Hybrid Clustering Method based on Improved Cuckoo Optimization and Modified Particle Swarm Optimization Algorithms.
2. Santosh Kumar Majhi& Shubhra Biswal 2018. Optimal cluster analysis using hybrid K-Means and Ant Lion Optimizer.
3. Xiaojuan Hu et.al 2017, MapReduce-based improvement algorithm for DBSCAN.

4. Chun Guan et.al 2019, Particle swarm Optimized Density-based Clustering and Classification: Supervised and unsupervised learning approaches.
5. Nabiha Azizi1 et.al 2017, Particle Swarm Optimization Based Swarm Intelligence for Active Learning Improvement: Application on Medical Data Classification.
6. Razieh Sheikhpour et.al 2015. Particle swarm optimization for bandwidth determination and feature selection of kernel density estimation based classifiers in diagnosis of breast cancer.
7. Vijayalakshmi et.al 2020. Multi-modal prediction of breast cancer using particle swarm optimization with non-dominating sorting.
8. Uzma et.al 2020. A feature selection technique through unsupervised deep learning-based clustering for large gene expression data.
9. Chun Guan2018, Evolutionary and Swarm Algorithm Optimized DensityBased Clustering and Classification for Data Analytics.
10. Limin Wang 2018, an improved density-based spatial clustering of application with noise.
11. ERICH SCHUBERT et.al 2017, DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN.
12. Rui Xu 2005, Clustering Algorithms.
13. Megha Mane 2013, Data mining using Association rule based on APRIORI algorithm and improved Approach with illustration.
14. Grace L. Samson Ph.D 2017. MINING COMPLEX SPATIAL PATTERNS.
15. Krzysztof Koperski et.al 2014. Spatial Data Mining: Progress and Challenges.
16. Tapas Ranjan Baitharu 2015.A Comparative Study of Data Mining Classification Techniques using Lung Cancer Data.
17. T. Soni Madhulatha 2012. AN OVERVIEW ON CLUSTERING METHODS.
18. Tom Chiu et.al 2014.A robust and scalable clustering algorithm for mixed type attributes in large database environment.
19. K. Sumathi et.al 2016. Data Mining: Analysis of student database using Classification Techniques.
20. Jason Brownlee 2020. Evaluate the Performance of Machine Learning Algorithms in Python using Resampling.
21. K. Mumtaz 2012. An Analysis on Density Based Clustering of Multi-Dimensional Spatial Data.
22. Surbhi Sharma 2017. Enhancing DBSCAN Algorithm for Data Mining.
23. ERICH SCHUBERT et.al 2017. DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN.
24. Divya; Insha Altaf 2017. Cluster analysis using gene expression data.