Journal of Education and Science (ISSN 1812-125X), Vol: 31, No: 02, 2022 (69-82)

A Review of Clustering Methods Based on Artificial Intelligent Techniques

Baydaa I. Khaleel

Computer Science Department, College of Computer Science and Mathematics, Mosul University, Mosul, IRAQ

E-mail: baydaaibraheem@uomosul.edu.iq

(Received February 21, 2022; Accepted April 21, 2022; Available online June 01, 2022)

DOI: 10.33899/edusj.2022.133092.1218, © 2022, College of Education for Pure Science, University of Mosul.

Abstract: Due to the development in various areas of life, the development of the Internet, and the presence of many datasets, and in order to obtain useful information from the rapidly increasing volumes of digital data, there must be theories and computational tools to help humans extract the useful information they need from this data. Large data is collected from many different services and resources. Clustering is one of the most basic and well-known methods of data mining and extraction and obtaining useful information. The technique of recognizing natural groups or clusters within several datasets based on some measure of similarity is known as data clustering. Many researchers have introduced and developed many clustering algorithms based on the different methods of artificial intelligence techniques. Finding the right algorithms greatly helps in organizing information and extracting the correct answer from different database queries. This paper provides an overview of the different clustering methods using artificial intelligence and finding the appropriate clustering algorithm to process different data sets. We highlight the best-performing clustering algorithm that gives effective and correct clustering for each data set.

Keywords: Clustering, Swarm Intelligence Techniques, Artificial Neural Networks, Fuzzy Clustering.

مراجعة لطرق العنقدة المعتمدة على تقنيات الذكاء الاصطناعي

بيداء ابراهيم خليل

قسم علوم الحاسوب، كلية علوم الحاسوب والرياضيات، جامعة الموصل، الموصل، العراق

الخلاصة

نظرا للتطور الحاصل في مجالات الحياة المختلفة وتطور الانترنت ووجود العديد من مجموعة البيانات الضخمة والكبيرة. و للحصول على المعلومات المفيدة من الاحجام المتزايدة بسرعة كبيرة من البيانات الرقمية لابد من وجود نظريات وادوات حسابية لمساعدة البشر لاستخراج المعلومة المفيدة والتي يحتاجها من بين هذه البيانات. والبيانات الكبيرة يتم جمعها من العديد من الخدمات والموارد المختلفة. والتجميع هو من اهم الطرق الاساسية المعروفة للتنقيب عن البيانات واستخراجها والحصول على المعلومة المفيدة. وتجميع البيانات هو عملية تحديد التجمعات الطبيعية او العناقيد ضمن البيانات المتعددة بناء على بعض مقاييس التشابه. قدم العديد من

الباحثين وطوروا العديد من خوارزميات التجميع المعتمدة على طرائق التقنيات الذكائية الاصطناعية المختلفة. ويساعد العثور على الخوارزميات المناسبة بشكل كبير على تنظيم المعلومات واستخراج الاجابة الصحيحة من الاستعلامات المختلفة لقواعد البيانات. وهذه الورقة تقدم لمحة عامة عن طرق التجميع المختلفة باستخدام طرق الذكاء الاصطناعي وايجاد خوارزمية التجميع المناسبة لمعالجة مجموعة البيانات المختلفة. وقمنا بتسليط الضوء على خوارزمية التجميع الافضل اداء والتي تمنح المجموعات الفعالة والصحيحة لكل مجموعة البيانات.

الكلمات المفتاحية: العنقدة، تقنيات ذكاء السرب، الشبكات العصبية الاصطناعية، العنقدة المضببة.

1. Introduction

Data clustering is a technique for detecting natural groupings or clusters within multidimensional data based on a similarity measure (for example Euclidean distance). In pattern recognition and machine learning, it's a crucial step. Furthermore, in Artificial Intelligence, data clustering is a crucial procedure (AI). Data mining, vector and color image quantization, image segmentation, compression machine learning, and other applications use clustering techniques [1][2]. A cluster center, also known as a centroid, is used to identify it. As data clusters can be of varied shapes and sizes, data clustering is a difficult task in unsupervised pattern identification [3][4]. A cluster is a group of similar features, and those features are not alike and from different groups. Fuzzy clustering methods work to assign each pattern to each group with a certain degree of membership [5]. In the current digital age, according to the tremendous progress and development of the Internet and technologies of the online world, and the data generated by machines and devices, the systems have reached a huge size that increases day by day and is expected to increase in the coming years. There is information that is not known in advance and may be useful, so to extract this implicit information, we use data mining and its clustering process that lies at the intersection of artificial intelligence, machine learning, statistics and database systems [6]. One of the most important tasks in data mining is cluster analysis, discovery and analysis, and is used in various applications such as speech processing, image processing, web applications, and information retrieval. Clustering basically aims to group the data set into groups so that similar data are aggregated together in the same group while different data should belong to different groups [1][6]. Clustering is known as one of the most difficult tasks. Different algorithms developed by researchers over the years lead to different sets of data, even for the same algorithm, the choice of different parameters or the order of displaying data objects may significantly affect the final clustering sections. However, with the huge number of surveys and comparative studies related to clustering algorithms, exploration of the algorithm that aggregates the diverse, sparse data set is still an open problem. Therefore, the main objective of this paper is to provide comprehensive reviews of clustering algorithms that optimally aggregate diverse, sparse datasets [7][8]. To achieve the goal, we seek to define many artificial intelligence algorithms represented by artificial neural networks algorithms, as well as fuzzy clustering algorithms and swarm intelligence algorithms, and by analyzing and comparing the performance of these methods, we provide readers with the best of these methods in clustering the data set [8][9].

2. Clustering

Cluster analysis is one of the fruits of scientific development in the field of data tabulation and classification, and it is one of the common and useful techniques in the process of data mining in order to discover aggregates and identify important distributions and patterns in basic data. The beginning of the topic dates back to the first half of the twentieth century [10]. The general idea of a cluster is to divide a given set of data into groups (clusters) so that data points in one cluster are more similar to data points in other clusters. To put it another way, data clustering is the process of

grouping data into related categories. The clustering method separates a collection of data into groupings. The similarity between two points in the same group is greater than the similarity between two points in different groups [11].

Previously, classification methods depended on great personal effort and personal scientific experience in a particular field in order to lead to accurate and reliable results, but the computer's invasion of broad fields of science made the process of controlling abundant and wide information easy, with little effort and little cost. This has prompted taxonomists in various sciences to turn to this great device to exploit its mighty capabilities in the interest of their specializations. The concept of data collection is a simple one in nature, and is extremely similar to how humans think, whenever a large amount of data is dealt with, to make the analysis process easier, it is common to group or categorize a large amount of data into a few groups or categories. Clustering methods are commonly used not just to organize and classify data, but also for data compression and data model creation. These algorithms are called cluster analysis [11][12].

2.1 Cluster Analysis

Cluster analysis was given this name because its results appeared in the form of clusters or aggregates, so that the elements of each cluster were similar to each other, that is similarity within the cluster and different from the rest of the clusters, meaning less similarity between the elements for the different clusters. It can be defined as a programmed multivariate statistical analysis process that depends on calculating a variety of variables, such as characteristics, properties, components, and so on, of many different models according to the particular issue, then comparing those models with each other depending on the variables they contain, and ordering Links to each other in the form of clusters [13]. These correlations can be explained to find out the relationships between those models, or finding similarities between the different variables after comparing them to each other and arranging them in clusters depending on their availability in the different models. Where in the first case pairs of models are compared to each other for all variables, while in the second case pairs of variables are compared to each other for all models.

Clustering can go under different names in different contexts, and used in many applications including engineering, medicine, electronics, anthropology, archeology, earth science, marketing and economics [14][15][16].

2.2 Development of the Clustering Process

The clustering process can be in the form of different aggregates resulting from the segmentation of the main data set depending on a special scale used in the cluster, so there is a need for preprocessing the data before applying the cluster to it. Figure (1) explains the basic steps for developing the clustering process, and can be explained in the following points [17][18]]19][20]:

Feature Selection: selecting attributes step, clustering is implemented that enables the analysis of more information codes related to the problem to be solved. Therefore, it is necessary to pre-process the data before using it in the clustering phase.

Clustering Algorithm Selection: the appropriate algorithm is selected that results from defining a good cluster schema for the data set. Cluster criterion and approximate scaling fundamentally characterize the clustering algorithm in addition to its efficiency in deciding which cluster schema best fits the data set.

Validity of Results: The results validity of the clustering algorithm is achieved by using criteria and appropriate techniques, and because clustering algorithms know the previously unknown clusters, regardless of the clustering algorithms, so the final segmentation of the data requires some types of evaluation in some cluster applications.

Results Interpretation: In most cases, experience in the field of application is needeeds to integrate clustering results with the rest of the demonstrated and analyzed experiments in order to get the correct conclusion [21][22][23].

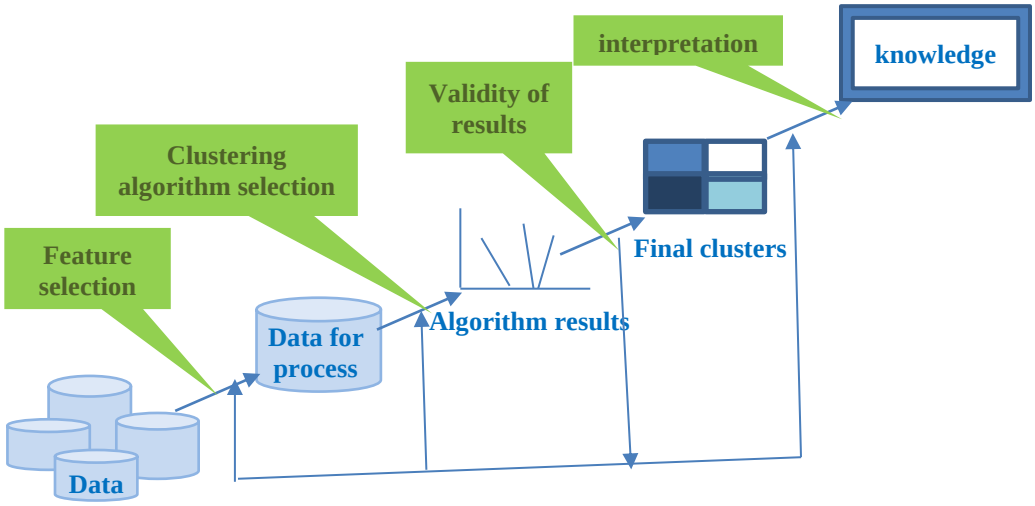

Figure 1: Knowledge discovery Process

3. Similarity and Distance Measures

There are many important characteristics in the formation of clusters by solving the specific cluster problem, and the most important of these characteristics is that the pair within one cluster is similar to the pairs within that cluster (similarity within the cluster) rather than being similar to the pairs outside that cluster (extra-cluster resemblance). The similarity scale aims to measure the closeness and similarity between the items with the highest similarity value. One of the characteristics required for a good similarity measure is that the percentage of similarity of any pair with itself is perfect, that is 100%, and that the percentage of similarity of any completely dissimilar pair is 0%, meaning that for every pair of pairs in the database, the first pair is not similar to the second pair. So the percentage of similarity between any two spouses is 100%, so that the first pair is completely similar to the second pair, and the percentage of similarity between any two pairs that have some similarity lies between zero and 100%. The cluster problem for any two pairs of data determines the affiliation of each pair to one of the group of clusters that have the data classified within a particular problem. Using the idea of similarity measures is not easy in the clustering process. So it is often used in place of measures of similarity, the measure of dissimilarity, or Distance (dis). The distance scale is the opposite of similarity, denoted by the symbol dis (p_i,p_t) and is used in clustering most often. So, suppose we have a cluster C_j, $1 \leq j \leq K$, K is the number of clusters, so that

$$dis(p_{jt}, p_{jn}) \leq dis(p_{jt}, p_i)\ \forall p_{jt}, p_{jn} \in C_j \text{ and } p_i! \in C_j$$

Assuming that we have the two vectors $p_i=\{p_{i1},\ldots, p_{io}\}$ and $p_j=\{p_{jl},\ldots, p_{jo}\}$ there are traditional distance measures that can be used in a two-dimensional space, and these distances are the Euclidean distance and the Manhattan distance as follows:

$$\text{Euclidean}: dis(p_i, p_j) = \sqrt{\sum_{t=1}^{o}(p_{it} - p_{jt})^2} \quad (1)$$

$$\text{Manhattan}: dis(p_i, p_j) = \sum_{t=1}^{o}|(p_{it} - p_{jt})| \quad (2)$$

The Euclidean distance depends on the distance between the vectors p_i and p_j, which is calculated as the square root of the sum of the squares of the arithmetic differences of the coordinates corresponding to two points [23].

4. Clustering in artificial intelligent techniques

In order to set the clustering algorithm in artificial neural networks and fuzzy clustering, it is necessary to determine how to take advantage of the parameters of each of these methods, because clustering algorithms are designed to find clusters that fit some patterns and depend on the correct selection of parameters. Swarm Intelligence Techniques are intelligent technique that involves

studying the collective behavior of decentralized systems. These techniques simulates the social behavior of insects or swarms in order to simplify the design of solutions to complex problems. Swarm intelligence has been successfully applied to the fields of optimization, military applications and robotics over the past few years [24][25].

The behavior of different animals can be observed in nature, for example, ants can find the shortest way to obtain food and assign workers to various tasks and defend their colonies, as well as a group of fish swimming in perfect harmony or a group of birds flying regularly and changing directions of flight without collidinge with each other. These groups have strong capabilities to solve complex problems, where the swarm's intelligence solves complex problems by a number of individuals without central control or support from a general model, the local interactions between individuals and their surrounding environment that generate the general behavior to solve these complex problems. Contingency strategy and high distribution of control are the two most important characteristics of swarm intelligence as it produces an autonomous system that is independent, robust, flexible, adaptable, scalable, affordable, and self-regulating [26][27].

In general, all methods of swarm intelligence depend on the community, where in this community there is a set of potential solutions. The ideal solution is searched for through iterative steps, and the members of the community change their positions within the search space through vector and non-vector communications as well as mating and genetic mutation in the evolutionary computation. There are two popular swarm intelligence theories that have been inspired by computational intelligence [28][29][30]:

1- Ant colony optimization ACO algorithm: this method has been successfully applied to many separate optimization problems and it mimics the behavior of ants [31].
2- The particle swarm optimization (PSO) algorithm: This algorithm simulates particle swarm behavior and was ideal for solving nonlinear problems of optimization with some limitations[32][33][34].

The basic principles of collective behavior of birds represent:

- Homogeneity: all particles have the same behavior model and the flock moves without a leader.
- Locality: the movement of each particle will be affected only by the movement of the neighboring particles, and vision is the most important sense for the particle to organize the flock.
- Collision Avoidance: The particle avoids its mates in the flock.
- Flock Centering: The particle tries to stay close to its fellow flock.
- Velocity Matching: The particle tries to keep pace with the speed of its colleagues in the flock.

In this article, some of the most important clustering methods based on artificial intelligence techniques have been reviewed. These methods are used in several fields and applications, and they have advantages and disadvantages. The following are some of these artificial intelligence techniques.

I) Clustering Technique Based on Partition

The basic idea of this type of partition-based clustering algorithm is that it considers the center of the data points as the corresponding center of mass. One of the most popular algorithms for this type of aggregation is K-means. The main idea of K-means is updating the cluster center, which is represented by data points center. The repetitive process continues according to the iterative computation until some convergence criteria are met. The evaluation measurement of this method: The result of the clustering is based on the average distance between any data point in the cluster and the cluster center for the same cluster and the average distance between different clusters. One of the advantages of this algorithm in general is that it has superior computational efficiency, but its time complexity is relatively low. The disadvantages of this type of clustering algorithms are that they are not suitable for non-convex data, are somewhat sensitive to outliers, and are easy to make at the local

optimum level, the number of clusters for data is preset, and for the number of clusters the clustering results are sensitive. The main application implemented by a K- means algorithm is clustering, which clusters the data and makes it into several clusters based on the input data.

II) Clustering Technique in Fuzzy Theory

The basic idea of clustering algorithms based on Fuzzy theory is that the discrete value of the belonging label {1.0} is changed to the continuous period [1.0] in order to describe the belonging relationship more logically. The basic idea of work FCM algorithms is that each data point in any cluster gets membership by improving object function. This type of the clustering includes several typical algorithms such as: FCM, BCFCM, PFCM, MDFCM, SFCM, FLICM, FCMS1, FCMS2, RFCMK, TEFCM, WIPFCM, and KWFLICM. Advantages: The probability of belonging to this type of cluster is more realistic, and in clustering the accuracy is relatively high, for real-world data it gives a very useful result, this type of algorithm is used when the obtained region of interest is insufficient or when the size of clusters The data is small. The main application implemented by these algorithms is to cluster the input data and divide it into several clusters. Each cluster contains similar data and differs from the data of other clusters. This algorithm can work as a classifier.

III) Clustering Technique by using Swarm Intelligence

The basic idea of swarm intelligence-based clustering algorithms is to simulate natural or biological behavior. The algorithms include both of PSO, ACO, ABC. The principle of an ants algorithm: is initially data is distributed randomly and based on an ants' decision to data select or not - for further operation. This process is repeated until a satisfactory result is obtained. The idea of PSO algorithm, at first, an initial population of particles is selected randomly, and particle clusters are then updated depending on the clusters center, speed and location of each particle, until a satisfactory clustering result is obtained. The work of algorithm ABC is the simulation of bee behavior in searching for food and determining the source of food, identifying the population of bees and sharing local and global information. The most important benefits of this clustering using these methods is to obtain the best global algorithm as well as being easy to understand, and it can be made at the optimal local level. As for its disadvantages, it is not suitable for high-dimensional data or large-sized data, as well as its low scalability and low operating efficiency as well. These algorithms implement several applications, and their main goal is to find the optimal solution. It can be among its applications to work as a classifier such as the Gray Wolves algorithm. It is also possible to hybridize these methods with other traditional methods to implement many tasks such as image clustering, pattern recognition, image retrieval.

IV) Model-Based Clustering Technique

The basic idea of the model-based Clustering algorithm is to match the acquired data with mathematical models, and this type of Clustering method can reveal the details of the features of each set, where each set represents a class or concept. One of the most widely used methods of this type is artificial neural networks such as BP, SOM, RBF. The nerve network consists of an input layer that contains a set of nodes that are used for data entry and the input layer is associated with a hidden layer or the output layer. Each node in the input layer is associated with the output layer nodes by weights. Through learning rules for each network and modifying weights, the output is calculated for each node in the output layer. One of the advantages of using these clustering methods is that the advanced and diverse models will provide good means to describe the data in an understandable and appropriate manner. Each neural network model has its own characteristics, which achieve satisfactory results in many areas. As for its disadvantages, it takes a relatively long time to obtain results, and the assumption in building the neural network structure is incorrect and the assembly results are sensitive to some models. Some application of neural networks are image and pattern recognition, binary and multi classification, and diagnostics.

V) Clustering Methods based on Evolutionary Methods

The most famous Evolutionary methods are evolutionary programming EP and genetic algorithm GA. Solutions from the current generation to the next generation are selected depending on its fitness. The most important advantages of this method in the assembly process are to support multi-objective optimization. It is effective for complex, large and undefined search spaces, and can protect itself from falling into the optimal local level. The disadvantages are that it does not deal with dynamic data, it needs complex operators to not go to the local optimum level, and it is not suitable for problems that contain determinants. GA is used for data clustering, feature selection, prognostic.

5. Application of Artificial Intelligent Techniques in Data Clustering

In many areas, information retrieval and text mining are needed, and for this process to be done efficiently, clustering is used to collect documents. The data is divided into a specific number of groups, so that each group is data that is similar to each other and different from the other group, meaning that the groups are not overlapping.

Researchers S. A. Abdulrahman, M. Roushdy, A. B. Salem used the aggregation method to obtain the similarity between the different data by using the Euclidean distance scale. The noise was removed from the data using the K-Means algorithm. To get the most useful and optimal features, they used GA, and to classify the data they used SVM. The proposed method, which combined these three methods, obtained high data clustering and classification accuracy. The proposed method was applied to three data sets (wine, abalone, network). To evaluate the efficiency this method was use three measures are Sensitivity, Specificity, and Positively Predicted [35].

$$Sensitivity(\%) = \frac{TP}{TP+FN} \times 100 \quad (3)$$

$$Specificity(\%) = \frac{TN}{TN+FP} \times 100 \quad (4)$$

$$Positively\ Predicted(\%) = \frac{TP}{TP+FP} \times 100 \quad (5)$$

Farmani proposed a new and efficient clustering method for data clustering by studying the characteristics of the k-means method and the ABC method, focusing on the strengths between them, and then linking these two algorithms to obtain the new clustering method. Each cluster centers will represent the food source in the research environment, and one of the possible solutions will represent the source of food. Here, the following function was used as a criterion for the efficient cluster [36].

$$E = \sum_{j=1}^{K} \sum_{Zj \in Cj} \|Zj - Cj\|^2 \quad (6)$$

Zhang and Ma proposed a new method to conduct clustering and achieve a better balance between exploitation and exploration, resulting in non-falling to the local optimum level. Where they used the ELPSO method, which has excellent convergence ability, it was hybridized with the fuzzy clustering algorithm represented by the FCM algorithm to form a new hybrid method, which is FCM-ELPSO. The proposed hybrid method has the ability to correct the cluster direction during the training process, which leads to obtaining the best results. Experimental experiments showed that the proposed method works well on the UCI dataset. The measure of evaluating clustering in this hybrid method is a cluster index [37].

Lakshmi Patibandla et al. proposed a method for clustering medical data, where one of the swarm intelligence algorithms PSO and the genetic algorithm (GA) were used to cluster the data efficiently and quickly. Where the two algorithms were run both, and the parameters of the two algorithms were changed in concert, and the implementation alternated between them according to the values of the double-effect factors for the hybridization estimation, to give the indication of the implementation of either the particle swarm optimization or the genetic algorithm. Where the proposed method was applied to a variety of medical information, and these were stable clinical data, namely, iris, heart, thyroid, diabetes, and chest risk, and the results were good for a variety of clusters and a number of different characteristics. Accuracy rate is the efficiency measure of the method [38].

A. Abubaker, A. Baharum, and M. Alrefaei proposed a new clustering method based on simulated annealing SA and multi-objective particle swarm optimization MOPSO. The new proposed method

is capable of automatic clustering, which is suitable for dividing the data sets into an appropriate number of clusters. A novel method MOPSOSA is combined between simulated annealing method and multi-objective particle swarm optimization method. In this paper, three indicators were improved at the same time to determine the optimal number of clusters for the data set. The first cluster validity index depends on the Euclidean distance, the second depends on the point symmetry distance, and the third depends on the short distance. The F-measure is the criterion of evaluating performance of clustering:

$$F(T,C) = \sum_{i=1}^{K_T} \frac{|T_i|}{n} \max_{Cj \in C} \{F(T_i, C_j)\} \quad (7)$$

where n is the dataset's number. The best value of F(T,C) is 1, and higher values of F(T,C) are better values [39].

Torres et al. presented a new hybrid method for determining the physical fitness of schoolchildren, using information from a database of 1,813 children of both sexes. Between 6-12 years old. The information or features that were used to know the child's physical fitness are horizontal jumping, flexibility and agility, where in the beginning the adaptive fuzzy grid inference system (ANFIS) was used, and then it was improved using the swarm intelligence algorithm represented by the particle optimization method PSO that was used to cluster and classify the data. The tests showed that the results obtained from the proposed method were good. Mean squared error (MSE) and Root mean squared error (RMSE) wear used to assess the approaches' performance [40].

Thrun and Ultsch, used Databionic swarm (DBS) clustering. DBS has high strength when there are widely different types of cluster structures depending on density and distance. It uses swarm intelligence technology; DBS aggregation technique is very powerful and robust with respect to outliers than traditional algorithms. DBS algorithms can be combined for aggregation and can be used for knowledge discovery because there is no prior knowledge of the required data[41].

Chan, Ke, and Im , presented a probability *k*-means clustering algorithm that was implemented with a neural network architecture for finding the clustering center located objects. They used two datasets : Dataset A1, Dataset A 2, Dataset A 3, 2-dimensional data sets with varying numbers of circular clusters. These data sets have 3000, 5250 and 7500 vectors scattered across 20, 35 and 50 predefined clusters, the accuracy rate is a measure of the method's efficiency[42].

Xu et al. introduced a new method called the sparse and robust fuzzy K-Means algorithm, to obtain more accurate clustering results. This proposed method reduced the objective function in order to facilitate dealing with the effect of outlier's values, considering the sparse membership values from the method of re-weighting, which is a sum of fuzzy K-Means weights with strong norms, and they used three benchmark two datasets are image datasets and rest is the MNIST3 database of handwritten digits. Two measures were used to evaluate clustering methods, Accuracy (ACC) and Normalized Mutual Information (NMI) [43].

Sun et al. introduced a new method for identifying cognitive interactions, called epiACO, which is based on an ant colony ACO improvement algorithm. The most notable features of epiACO are path selection strategies and fitness function as well as a memory-based strategy. The value of the fitness function takes advantage of both the Bayesian network and the mutual information to measure associations efficiently and effectively between phenotype and SNPs. This method provides two strategies for path selection. One is to choose the random path and the other is to choose the probabilistic path, so as to guide the behavior of ants in an adaptive way for exploration and exploitation. The memory-based strategy is designed to retain candidate solutions obtained in previous iterations. These candidate solutions are compared with the solutions obtained from the current iteration to generate new candidate solutions. Using this method, the knowledge was determined with high accuracy. And the evaluation measure of identifying epistasis is Detection power that equal:

Power = R/ G

Where R is the number of datasets in which epistasis models have been successfully detected, and G is the number of all data sets [44].

Mortezanezhad and Daneshifar used the development algorithm for the process of clustering big data arising from the Internet of Things (IoT). They proposed a new clustering algorithm based on the GA genetic algorithm, which is unsupervised, that is, it does not require prior knowledge of the number of clusters and was used to classify the data into several groups. The proposed method used crossover operators and related mutations as well as a very short chromosome encoding and this led to a very good clustering performance. The data set used is balanced and unbalanced real world data which consists of 13 data vectors, as well as the use of random data consisting of one million samples that artificially generated. The proposed algorithm reached a high efficiency in the clustering process. The best fitness function and the shortest Euclidean distance were calculated as measures of quality and efficiency of the clustering method used [45].

Choudhry and Kapoor suggested different versions of FCM algorithm such as BCFCM, PFCM, MDFCM, SFCM, FLICM, FCMS1, FCMS2, RFCMK, TEFCM, WIPFCM, and KWFLICM. This is done by using the spatial statistics in the images. Depending on the performance of each method applied to noisy MRI images, all these segmentation algorithms are applied for two types of cluster validity functions namely, feature structure and fuzzy partition. In this paper, the partition coefficient, time complexity, partition entropy, and segmentation precision are evaluated for these variations in the FCM method. To measure the efficiency of the method, Segmentation Accuracy (SA) was used [46].

$$SA = \frac{Number\ of\ Correctly\ Classified\ Pixels}{Total\ Number\ of\ Pixels} \quad (8)$$

Elhoseny et al. introduced a new clustering and routing protocol to improve the work of the wireless sensor network, where swarm intelligence algorithms were used for this purpose. The particle swarm optimization algorithm was used to select and organize the group heads with high efficiency in time by clustering process, then the gray wolf optimization algorithm was used to conduct the routing process in order to determine the optimal paths in the network. This proposed IPSO-GWO method gave excellent results as it provided the greatest possible energy efficiency with a lifetime of the network. To determine the efficiency of this strategy, Energy efficiency measure was used to calculate the amount of energy used by each node over the course of a given time period [47].

Hussen, used the Backpropagation network and the K-Means algorithm to cluster groups of diverse medical data. These two methods were applied to three data sets for different diseases. The data set is for (heart disease, breast cancer and diabetes) and it has a high classification and aggregation performance. The accuracy rate is a measure of the method's efficiency [48].

Semchedine and moussaoui suggested a new fuzzy CMeans Algorithm (FCM) initialization strategy based on fuzzy particle swarm optimization (FPSO) which was applied to segmentation of images of the brain by magnetic resonance MR. In order to obtain the initial cluster centers for the FCM algorithm, the new proposed method called FPSOFCM used FPSO algorithm for this purpose, depending on a new fitness function that integrates the validity indicators of the fuzzy cluster, the criteria used to evaluate clustering was partition entropy (PE)

$$PE = \frac{-\sum_{i=1}^{N}\sum_{k=1}^{c} u_{ki}\, log(u_{ki})}{N} \quad (9)$$

When the value of (PE) is minimal, the best clustering is achieved [49].

Zhu et al. proposed an automatic clustering method called PSO-CFDP for subclassification of cancer. By using and applying the PSO particle swarm optimization algorithm several times, the cut-off distance is determined and the centers of mass are determined automatically. Experiments were performed on four standard data sets and two real cancer gene expression data sets. This method was able in a short time and cost to determine the centers of mass and the cutting distance automatically and got a good classification accuracy. The evaluative measure here is Accuracy rate [50].

Hoomod and Jebur, used Radial-Basis Function Networks (RBFN) and Self-Organizing Map (SOM) to improve Mobiles Ad-Hoc Networks (MANET) by dividing the network into different groups

using self-organization map SOM neural network. They also Radial Based Neural Network to increase the life time of nodes, packet delivery ratio because the optimal path has the best parameters from other paths. And the dataset is MRBFNN routing (fixed 10 nodes 2 clusters), MRBFNN routing (fixed 30 nodes 2 clusters). To evaluate the performance of these methods, the Euclidean distance to reach the optimal path between nodes of wireless network was calculate [51].

Farmani, used the Ant Colony algorithm (ACO) to perform the clustering process for the unlabeled data, as it relied on the automatic clustering of large sets of unlabeled data in advance. The method proposed by the researcher uses the principle of opposition-based learning while creating or configuring ant hunting sites in the API. The proposed improved method is called opposition based API. The results of the proposed method were compared with the traditional method for clustering and three other methods, and it became clear that the proposed algorithm was the best in reaching the optimal number of clusters and the data was excellently classified. Accuracy rate was the measure to evaluate this method [36].

Table 1 summarizes all these methodologies with different parameters like datasets used, and researchers, for easy and quick reference.

Table 1. Various artificial intelligent techniques based data clustering methods

Paper referred	Clustering Algorithm	Dataset
S. A. Abdulrahman et al.[35]	K-Means , GA, SVM	data sets (wine, abalone, network)
M. R. Farmani [36]	k-means method and the ABC method. Ant Colony algorithm (ACO), the principle of opposition-based learning while creating or configuring ant hunting sites in the API	pilot datasets These are 2- and higher-dimensional datasets with outliers and intra- and inter-clusters variations, which consist of different shapes of clusters (such as spiral, circular, elongated)
J. Zhang and Z. Ma [37]	FCM-ELPSO	Various datasets Abalone, Ecoli, Glass, Spectf, Steel plates faults, Ultrasonic flowmeter diagnostics, Yeast
R. S. M. Lakshmi Patibandla et al.[38]	PSO and the genetic algorithm GA	Medical dataset Iris, Breast cancer, Diabetes, Heart, Throid, Pima
A. Abubaker, A. Baharum, and M. Alrefaei [39]	Swarm optimization method MOPSO, SA, and hybrids method MOPSOSA	Real Life and Artificial datasets
J. S. Torres et al. [40]	adaptive fuzzy grid inference system (ANFIS)	database of 1813 records, MEN AND WOMEN FROM 6 TO 12 YEARS OLD, FROM AREQUIPA, PERU.
M.C.Thrun and A. Ultsch [41]	Databionic swarm (DBS) clustering	data about Covid-19 pandemic
K. Chan, W. Ke, and S. Im [42]	*k*-means clustering algorithm with a neural network architecture	Dataset S1, Dataset S2, Dataset S3, Dataset S4, These datasets contains from 5000 vectors and 15 clusters.
J. Xu, J. Han, K. Xiong, and F. Nie [43]	fuzzy K-Means with robust norms	DATASETS: COIL-20] Samples 1440 Dimensions 60 Classes 20, COIL-100 Samples 7200 Dimensions 160 Classes 100 MINIST-2K2K Samples 4000 Dimensions 120 Classes 10 MINIST-10K Samples 10000 Dimensions 120 Classes 10 MINIST-TEST Samples 10000 Dimensions 115 Classes 10 MINIST-ORIG Samples

		70000 Dimensions 120 Classes 10
Y. Sun, J. Shang et al. [44]	epiACO, based on an ant colony ACO improvement algorithm	A real data set of age-related macular
A. Mortezanezhad, and E. Daneshifar [45]	Proposed GA genetic algorithm	Balanced and unbalanced real world data
M. S. Choudhry and R. Kapoor [46]	Various FCM versions, such as BCFCM, PFCM, FLICM, SFCM, MDFCM, FCMS1, FCMS2, TEFCM, WIPFCM , RFCMK, and KWFLICM	medical images, MRI medical images
M. Elhoseny et al.[47]	proposed IPSO-GWO method	dataset for wireless sensor network
E. A. Hussen [48]	backpropagation network and the K-Means	dataset (heart disease, breast cancer and diabetes)
M. Semchedine and A. moussaoui [49]	FCM and (FPSO)	MRI Brain images
X. Zhu, J. Shang et al.[50]	Automatic clustering method called PSO-CFDP	Two real cancer gene expression data sets
H. K. Hoomod and T. K. Jebur [51]	Neural Network: self-organizing map (SOM), radial basis function RBF network	dataset is MRBFNN routing (fixed 10 nodes 2 clusters), MRBFNN routing (fixed 30 nodes 2 clusters)

Through these works, carried out by numerous researchers to cluster various sets of data using various intelligent methods, they reached satisfactory results to cluster their data and used various quality measures as well as various parameters according to the diversity of their methods they used. It is difficult to prefer one method over another and describe it as the best. This is perhaps the best for the data they used but for other datasets this method is perhaps not efficient, meaning that depending on the nature and quality of the data set used, it is possible to choose clustering methods that give promising results. We noticed that when using intelligent hybrid methods, the results were better.

6. Conclusion

The objective of this article is to conduct a detailed survey of different types of artificial clustering technologies. This paper reviewed methods for improving clustering using artificial intelligence techniques such as artificial neural network algorithms, genetic algorithms, and fuzzy clustering algorithms, as well as swarm optimization algorithms. Various measures were used to measure the performance efficiency of the clustering process. It is not possible to determine that any particular algorithm is good in all circumstances. Each algorithm has its strong advantages depending on the specific nature of the data, but for another type of data, this algorithm can fail and not reach a satisfactory result. Depending on the nature of the data, as well as making an appropriate decision about certain parameters, the selection of the method of clustering may be appropriate. When intelligent algorithms are hybridized among themselves or AI algorithms are hybridized with other algorithms, research reveals that they generate superior outcomes in many optimization issues in terms of efficiency and accuracy. When these hybrid algorithms for data clustering are used, the ideal number of clusters is produced, resulting in superior data prediction and analysis.

7. Acknowledgement

My thanks and gratitude to the University of Mosul/Iraq for its support in accomplishing this research.

8. References

[1] C. Benabdellah, A. Benghabrit, and I. Bouhaddou,"A survey of clustering algorithms for an industrial context", *Procedia Computer Science*, vol. 148, pp.291-302, 2019.

[2] J. Oyelade, I. Isewon, O. Oladipupo, O. Emebo, Z. Omogbadegun, O. Aromolaran, E. Uwoghiren, D. Olaniyan and O. Olawole," Data Clustering: Algorithms and Its Applications", *19th International Conference on Computational Science and Its Applications (ICCSA), IEEE*, pp.71-81, 2019, doi 10.1109/ICCSA.2019.000-1.

[3] S. Singh, and S. Srivastava, "Review of Clustering Techniques in Control System" *International Conference on Smart Sustainable Intelligent Computing and Applications under ICITETM2020, Procedia Computer Science*, 173, pp.272–280, 2020.

[4] N. Valarmathy, and S. Krishnaveni, "Performance Evaluation and Comparison of Clustering Algorithms used in Educational Data Mining", *International Journal of Recent Technology and Engineering (IJRTE)*, vol. 7, pp.103-113, 2019.

[5] M. G.H. Omran, A. P. Engelbrecht, and A. Salman,"An Overview of Clustering Methods", *Intelligent Data Analysis*, 2017, doi: 10.3233/IDA-2007-11602·

[6] I. Kholod, A. N. Rukavitsyn, N. Reva, and A. V. Shorov,"Distributed Data Clustering by Neural Network Algorithms", *IEEE*, pp.249-253, 2019, doi: 10.1109/EIConRus.2019.8657175.

[7] P. Radtha, and R. Divya, "An Efficient Detection of HCC-recurrence in Clinical Data Processing using Boosted Decision Tree Classifier", *International Conference on Computational Intelligence and Data Science, Procedia Computer Science*, 167, pp.193-204, 2020.

[8] K. Chen, A. Yadav, A. Khan, and K. Zhu, "Credit Fraud Detection Based on Hybrid Credit Scoring Model", *International Conference on Computational Intelligence and Data Science, Procedia Computer Science*, 167, pp.2-8, 2020.

[9] J. Serey, L. Quezada, M. Alfaro, G. Fuertes, M. Vargas, R. Ternero, J. Sabattin, C. Duran and S. Gutierrez, "Artificial Intelligence Methodologies for Data Management", *MDPI*, pp.1-30, 2021.

[10] M. Z. Rodriguez, C. H. Cominid, D. Casanova, O.M. Bruno, D.R. Amancio, L. F. Costa, and F. A. Rodrigues, "Clustering algorithms: A comparative approach", *POLS*, pp.1-34, 2019, doi.org/10.1371/journal.pone.0210236.

[11] M. Rahmani, E. Azhir, S. Ali, M. Mohammadi, O. H. Ahmed, M. Y. Ghafour, S. H. Ahmed and M. Hosseinzadeh, "Artificial intelligence approaches and mechanisms for big data analytics: a systematic study", *Peer Journal Computer Science*, pp.1-28, 2021.

[12] C. Ketels, "Cluster Mapping as a Tool for Development", *Institute for Strategy and Competitiveness Harvard Business School*, pp. 1-52, 2017.

[13] N. S. Rania, U. Karthik, and S. Ranjith, "Extraction of Gliomas from 3D MRI Images using Convolution Kernel Processing and Adaptive Thresholding", *Procedia Computer Science*, vol. 167, pp.273-284, 2020.

[14] Z. Liu, C. Ren and W. Cai, "Overview of clustering analysis algorithms in unknown protocol recognition", *MATEC Web of Conferences*, vol. 309, *EDP Sciences*, 2020, doi.org/10.1051/matecconf/202030903008.

[15] N. Derlukiewicz, A. M. zyk, D. Mankowska, A. Dyjakon, S. Minta and T. Pilawka, "How do Clusters Foster Sustainable Development? An Analysis of EU Policies", *MDPI*, pp. 1- 15, 2020, doi:10.3390/su12041297.

[16] N. Ruhil, M. Singh, D. Mitrac, A, Singh, and K. K. Singh, "Detection of changes from Satellite Images Using Fused Differene Images and Hybrid Kohonen Fuzzy C-Means Sigma", *International Conference on Computational Intelligence and Data Science*, vol. 167, pp.431-439, 2020.

[17] R. Kumar, and A. Singh, "Robustness in Multilayer Networks Under Strategical and Random Attackes", *Procedia Computer Science*, vol. 173, pp.94-103, 2020.

[18] K. Sumiran, "An Overview of Data Mining Techniques and Their Application in Industrial Engineering", *Asian Journal of Applied Science and Technology (AJAST),*) vol. 2, Issue 2, pp. 947-953, 2018.

[19] V. K. Chawra, and G. P. Gupta, "Load Balanced Node Clustering scheme using Improved Memetic Algorithm based Meta-heuristic Technique for Wireless Sensor Network" *International Conference on Computational Intelligence and Data Science (ICCIDS 2019), Procedia Computer Science*, vol. 167,pp.468-476, 2020.

[20] M. Faizan, M. F. Zuhairi, S. Ismail, and S. Sultan, "Applications of Clustering Techniques in Data Mining: A Comparative Study", *(IJACSA) International Journal of Advanced Computer Science and Applications*, vol. 11, no. 12, pp.146-153, 2020.

[21] C. Ketels, "Cluster Mapping as a Tool for Development", *Institute for Strategy and Competitiveness Harvard Business School*, pp.1-52, 2017.

[22] Y. I. Hamodi, R. R. Hussein, and N. T. Yousir, "Development of a Unifying Theory for Data Mining Using Clustering Techniques", *Webology*, vol. 17, no. 2, pp.1-13, 2020.

[23] D. S. Rohmah, S. D. R. Sari, and K. V. Yugi, "Clustering Human Development Index Data with Gravitational Search Algorithm-Fuzzy 4-Means (GSA-F4M)", *AIP Conference Proceedings*, pp.1-9, 2021.

[24] H. Sarker, "Machine Learning: Algorithms, Real-World Applications and Research Directions", *SN Computer Science*, pp. 2-160, 2021.

[25] L. Hoti, "Application of Artificial Intelligence Techniques to Combat Money Laundering in the Banking Sector", *master Thesis, Stockholm University*, 2021.

[26] N. Sharma, and V. Guptab, "Meta-heuristic based optimization of WSNs Localisation Problem- a Survey", International Conference on Smart Sustainable Intelligent Computing and Applications under ICITETM2020, vol. 173, pp.36-45, 2020.

[27] G. Kicska, and A.Kiss, "Comparing Swarm Intelligence Algorithms for Dimension Reduction in Machine Learning", *MDPI*, pp.1-15, 2021, doi.org/10.3390/bdcc5030036.

[28] J. Ning, C. Zhang, P. Sun and Y. Feng, "Comparative Study of Ant Colony Algorithms for Multi-Objective Optimization", *MDPI*, pp.1-19, 2018, doi:10.3390/info10010011.

[29] B. S. Harish, S. V. A. Kumar, F. Masulli and S. Rovetta, "Adaptive Initialization of Cluster Centers using Ant Colony Optimization: Application to Medical Images", *6th International Conference on Pattern Recognition Applications and Methods*, pp.591-598, 2017, doi: 10.5220/0006210905910598.

[30] S. Girsang, T. W. Cenggoro, and K. W. Huang, "Fast Ant Colony Optimization for Clustering", *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 12, no. 1, pp.78-86, 2018.

[31] S. Gupta, G. Gautam, Deepak, D. Vats, P. Varshney, and S. Srivastava, "Estimation of Parameters in Fractional order Financial Chaotic system with Nature Inspired Algorithms", International Conference on Smart Sustainable Intelligent Computing and Applications under ICITEM2020, vol. 173, pp.18-27, 2020.

[32] O. F. Aje, and A. A. Josephat, "The particle swarm optimization (PSO) algorithm application – A review", *Global Journal of Engineering and Technology Advances*, vol. 03, pp. 001-006, 2020.

[33] D. Wang, D. Ta, and L. Liu, "Particle swarm optimization algorithm: an overview", *Springer, Soft Computing*, 2017, doi 10.1007/s00500-016-2474-6.

[34] G. Rossides, B. Metcalfe, and A. Hunter, "Particle Swarm Optimization-An Adaptation for the Control of Robotic Swarms", *MDPI*, pp.1-21, 2021, doi.org/10.3390/robotics10020058.

[35] S. A. Abdulrahman, M. Roushdy, and A. B. Salem, "Intelligent Clustering Technique based on Genetic Algorithm", *International Journal of Intelligent Computing and Information Sciences*, vol.21, no.1, pp.19-32, 2021.

[36] M. R. Farmani," Clustering Analysis using Swarm Intelligence", *Ph.D. Thesis in Electronic and Computer Engineering*, University of Cagliari, 2015.

[37] J. Zhang and Z. Ma, "Hybrid Fuzzy Clustering Method Based on FCM and Enhanced Logarithmical PSO (ELPSO)", *Hindawi, Computational Intelligence and Neuroscience*, vol. 2020, Article ID 1386839,pp.1-12, 2020.

[38] R.S.M. L. Patibandla, B. T. Rao, P. S. Krishna, and V. R. Maddumala, "Medical Data Clustering using Particle Swarm Optimization Method", *Journal of Critical Reviews*, vol. 7, pp.363-367, 2020.

[39] A. Abubaker, A. Baharum, and M. Alrefaei," Automatic Clustering Using Multi-objective Particle Swarm and Simulated Annealing", *POLS*, pp.1-23, 2015, doi:10.1371/journal.pone.0130995.

[40] J. S. Torres, G. L. Luza, D. C. Yana, J. G. Valdivia, and M. C. Bolanos, "Neuro-fuzzy System with Particle Swarm Optimization for Classification of Physical Fitness in School Children", *(IJACSA) International Journal of Advanced Computer Science and Applications*, vol. 11, no. 6, pp.505-512, 2020.

[41] M. C. Thrun and A. Ultsch," Swarm Intelligence for Self-organized Clustering (Extended Abstract)", *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20) Journal Track*, pp.5125-5129, 2020.

[42] K. Chan, W. Ke, and S. Im," Probability *k*-means Clustering for Neural Network Architecture", *ICAAI, Istanbul, Turkey*, pp.52-57, 2021.

[43] J. Xu, J. Han, K. Xiong, and F. Nie, " Robust and Sparse Fuzzy K-Means Clustering", *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16)*, pp. 2224- 2230, 2016.

[44] Y. Sun, J. Shang , J. Liu, S. Li, and C. Zheng, "epiACO - a method for identifying epistasis based on ant Colony optimization algorithm", *National center for Biotechnology information*, vol. 10, pp.10-23, 2017, doi: 10.1186/s13040-017-0143-7.

[45] A. Mortezanezhad, and E. Daneshifar, "Big-Data Clustering with Genetic Algorithm", *5th Conference on Knowledge Based Engineering and Innovation (KBEI)*, 2019, https://ieeexplore.ieee.org/document/8735076.

[46] M. S. Choudhry and R. Kapoor," Performance Analysis of Fuzzy C-Means Clustering Methods for MRI Image Segmentation", *Twelfth International Multi-Conference on Information Processing (IMCIP), ELSEVIER, Procedia Computer Science*, vol. 89, pp.749-758, 2016.

[47] M. Elhoseny, R. S. Rajan, M. Hammoudeh, K. Shankar, and O. Aldabbas,"Swarm intelligence–based energy efficient clustering with multihop routing protocol for sustainable wireless sensor networks", *International Journal of Distributed Sensor Networks*, pp.1-12, 2020.

[48] S. A. Hussen, "Comparison of Performance Between Back Propagation and K-means on Medical Datasets", *Journal of Babylon University/Pure and Applied Sciences*, vol. 26, no. 3, pp.6-9, 2018.

[49] M. Semchedine and A. Moussaoui, "An Efficient Particle Swarm Optimization for MRI Fuzzy Segmentation", *Romanian Journal of Information Science and technology*, vol. 20, no. 3, pp. 271-285, 2017.

[50] X. Zhu, J. Shang, Y. Sun, F. Li, X. Liu, and S. Yuan, "PSO-CFDP: A Particle Swarm Optimization-Based Automatic Density Peaks Clustering Method for Cancer Subtyping", *Hum Hered*, vol. 84, pp.9-20, 2019, doi:10.1159/000501481.

[51] H. K. Hoomod, and T. K. Jebur, "Applying self-organizing map and modified radial based neural network for clustering and routing optimal path in wireless network", *IOP Conf. Series: Journal of Physics:Conf. Series 1003*, pp.1-11, 2018, doi :10.1088/1742-6596/1003/1/012040.