Database Clustering using Intelligent Techniques

Abbas Khudhair Abbas[*1] and Ahmed Tariq Sadeq[**2]
[*]Department of Computer and Educational Continues, Al-Nahrain University.
[**]Computer Science, Technology of University.
[1]E-mail: abbas_soc@yahoo.com.
[2]E-mail: Drahmed_Tarik@yahoo.com.

Abstract

Owning to the huge amounts of data collected in database, cluster analysis has recently become a highly active topic in data mining research. In data mining, efforts have focused on finding methods for efficient and effective cluster analysis in large database.

This paper proposes two new partitioning cluster methods, first is modified k-mean clustering algorithm with variable Neighborhood Search as a metaheuristic search and the second is modified k-mean clustering algorithm with cuckoo search as swarm intelligence.

The proposed algorithms does not need to enter the value of cluster points, instead of that it finds it automatically to get the best clustering using the clustering validity. This represents its fundamental characteristic.

The experiments were made on a many different sizes of databases some of the obtained from University of California (UC) Irvine Machine Learning Repository which maintain 246 data sets as a service to the machine learning community.

From these experiments, it is concluded that these methods reduced the time which needed to get the best solution as a half time which needed to perform same actions and in the same time it reduced the iterations to get the best solution. In addition, these proposed clustering methods give best quality (as performance) compared with other clustering methods; the performance was improved between (10% - 20%) compared with the original k-mean clustering method.

Introduction

The term "clustering" is used in several research communities to describe methods for grouping of unlabeled data. These communities have different terminologies and assumptions for the components of the clustering process and the contexts in which clustering are used. Thus, we face a dilemma regarding the scope of this survey. The production of a truly comprehensive survey would be a monumental task given the sheer mass of literature in this area. The accessibility of the survey might also be questionable given the need to reconcile very different vocabularies and assumptions regarding clustering in the various communities [1].

Clustering is the unsupervised classification of patterns (observations, data items, or feature vectors) into groups (clusters). The clustering problem has been addressed in many contexts and by researchers in many disciplines; this reflects its broad appeal and usefulness as one of the steps in exploratory data analysis. However, clustering is a difficult problem combinatorially, and differences in assumptions and contexts in different communities has made the transfer of useful generic concepts and methodologies slow to occur.

Clustering is useful in several exploratory pattern-analysis, grouping, decision making, and machine-learning situations; including data mining, document retrieval, image segmentation, and pattern classification. However, in many such problems, there is little prior information (e.g., statistical models) available about the data, and the decision-maker must make as few assumptions about the data as possible. It is under these restrictions that clustering methodology is particularly appropriate for the exploration of interrelationships among the data points to make an assessment (perhaps preliminary) of their structure.

Clustering Techniques

The main reason for having many clustering Techniques (methods) is the fact that the notion of "cluster" is not precisely defined (Estivill-Castro, 2000). Consequently

many clustering methods have been developed, each of which uses a different induction principle [2].

An important issue in clustering is how to determine the similarity between two objects, so that clusters can be formed from objects with high similarity within clusters and low similarity between clusters. Commonly, to measure similarity or dissimilarity between objects, a distance measure such as Euclidean, Manhattan and Minkowski is used. A distance function returns a lower value for pairs of objects that are more similar to one another [2].

Partitioning Methods

Partitioning methods relocate instances by moving them from one cluster to another, starting from an initial partitioning. Such methods typically require that the number of clusters will be pre-set by the user. To achieve global optimality in partitioned-based clustering, an exhaustive enumeration process of all possible partitions is required. Because this is not feasible, certain greedy heuristics are used in the form of iterative optimization. Namely, a relocation method iteratively relocates points between the *k* clusters. The following subsections present various types of partitioning methods.

There are many partition clustering techniques [2]:

- **K-Means**: is a commonly used algorithm. The aim of *K-Means* clustering is the optimization of an objective function that is described by the equation:

$$E = \sum_{i=1}^{C} \sum_{x \in C} d(x , m_i) \quad \text{.....(1)}$$

In the above equation, m_i is the center of cluster C_i , while $d(x,m_i)$ is the Euclidean distance between a point x and m_i . Thus, the criterion function E attempts to minimize the distance of each point from the center of the cluster to which the point belongs. The algorithm begins by initializing a set of *c* cluster centers.

- **PAM** (*Partitioning Around Medoids*). The objective of PAM is to determine a representative object (*medoid*) for each cluster, that is, to find the most centrally located objects within the clusters.
- **CLARA** (*Clustering Large Applications*), is an implementation of PAM in a subset of the dataset. It draws multiple samples of the dataset, applies PAM on samples, and then outputs the best clustering out of these samples.
- **CLARANS** (*Clustering Large Applications based on Randomized Search*), combines the sampling techniques with PAM. The clustering process can be presented as searching a graph where every node is a potential solution, that is, a set of *k*-medoids. The clustering obtained after replacing a medoid is called the *neighbour* of the current clustering. CLARANS selects a node and compare it with a user-defined number of their neighbours searching for a local minimum. If a better neighbour is found (i.e., having lower-square error), CLARANS moves to the neighbour's node and the process start again; otherwise the current clustering is a local optimum. If the local optimum is found, CLARANS starts with a new randomly selected node in search for a new local optimum [3].

K-Means Clustering

K-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori. The main idea is to define k centroids, one for each cluster. These centroids should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early group-age is done. At this point we need to re-calculate k new centroids as bary-enters of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid. A loop has been generated. As a result of this loop we may notice that the k centroids change their location step by step until no more changes are

done. In other words centroids do not move any more [4].

Finally, this algorithm aims at minimizing an *objective function*, in this case a squared error function. The objective function:

$$J = \sum_{j-1}^{x} \sum_{i-1}^{x} \left\| x_i^{(j)} - c_j \right\|^2 \quad \text{.....(2)}$$

Where $\left\| x_i^{(j)} - c_j \right\|^2$ is a chosen distance measureed between a data point $X_i^{(j)}$ and the cluster centre. C_j, is an indicator of the distance of the *n* data points from their respective cluster centres. The algorithm is composed of the following steps:

1. *Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.*
2. *Assign each object to the group that has the closest centroid.*
3. *When all objects have been assigned, recalculate the positions of the K centroids.*
4. *Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.*

Algorithm 1: K-Mean Algorithm

Although it can be proved that the procedure will always terminate, the k-means algorithm does not necessarily find the most optimal configuration, corresponding to the global objective function minimum. The algorithm is also significantly sensitive to the initial randomly selected cluster centers. The k-means algorithm can be run multiple times to reduce this effect [5].

Intelligent Technique

Artificial intelligence (AI) is technology and a branch of computer science that studies and develops intelligent machines and software. Major AI researchers and textbooks define the field as "the study and design of intelligent agents", where an intelligent agent is a system that perceives its environment and takes actions that maximize its chances of success. John McCarthy, who coined the term in 1955, defines it as "the science and engineering of making intelligent machines" [6].

Many problems in AI can be solved in theory by intelligently searching through many possible solutions: Reasoning can be reduced to performing a search. For example, logical proof can be viewed as searching for a path that leads from premises to conclusions, where each step is the application of an inference rule. Planning algorithms search through trees of goals and sub goals, attempting to find a path to a target goal, a process called means-ends analysis. Robotics algorithms for moving limbs and grasping objects use local searches in configuration space. Many learning algorithms use search algorithms based on optimization [7].

Metaheuristic and Swarm Intelligent

Computing optimal solutions is intractable for many optimization problems of industrial and scientific importance. In practice, we are usually satisfied with "good" solutions, which are obtained by heuristic or metaheuristic algorithms. Metaheuristics represent a family of approximate optimization techniques that gained a lot of popularity in the past two decades. They are among the most promising and successful techniques. Metaheuristics provide "acceptable" solutions in a reasonable time for solving hard and complex problems in science and engineering. This explains the significant growth of interest in metaheuristic domain. Unlike exact optimization algorithms, metaheuristics do not guarantee the optimality of the obtained solutions. Instead of approximation algorithms, metaheuristics do not define how close the obtained solutions from the optimal ones are.

Metaheuristics are a branch of optimization in computer science and applied mathematics that are related to algorithms and computational complexity theory. The past 20 years have witnessed the development of numerous metaheuristics in various communities that sit at the intersection of several fields, including artificial intelligence, computational intelligence, soft computing, mathematical programming, and operations research. Most of the metaheuristics mimic

natural metaphors to solve complex optimization problems (e.g., evolution of species, annealing process, ant colony, particle swarm, immune system, bee colony, and wasp swarm) [8].

Swarm intelligence is an important concept in artificial intelligence and computer science with emergent properties. The essential idea of swarm intelligence algorithms is to employ many simple agents applying almost no rule which in turn leads to an emergent global behavior.

To put it in a simple way, swarm intelligence can be described as the collective behavior emerged from social insects working under very few rules. Self-organization is the main theme with limited restrictions from interactions among agents. Many famous examples of swarm intelligence come from the world of animals, such as birds flock, fish school and bugs swarm [9].

Variable Neighborhood Search Algorithm

Variable Neighborhood Search is a Metaheuristic and a Global Optimization technique that manages a Local Search technique. It is related to the Iterative Local Search algorithm.

The strategy for the Variable Neighborhood Search involves iterative exploration of larger and larger neighborhoods for a given local optima until an improvement is located after which time the search across expanding neighborhoods is repeated. The strategy is motivated by three principles: [10]

1) a local minimum for one neighborhood structure may not be a local minimum for a different neighborhood structure,
2) a global minimum is a local minimum for all possible neighborhood structures, and
3) local minima are relatively close to global minima for many problem classes.

Variable Neighborhood Search (VNS) is a recent metaheuristic, or framework for building heuristics, which exploits systematically the idea of neighborhood change, both in the descent to local minima and in the escape from the valleys which contain them [11].

Modified K-mean with VNS Algorithm

In Variable Neighborhood Search, we used new approach to generate the value of clustering points automatically depending on minimum and maximum value in data set after enter the number of clusters which entered manually this divide the range into some categories its number equal to the number of clusters. After that we use the k-mean algorithm approach to find the closest value in each cluster using error squared function to force the fitness on the generated clusters to obtain the best solutions in minimum iteration and short time compared with original k-mean clustering algorithm. [12]

Initialization, select the clustering points value from dividing the range (max – min) on cluster number required ;
Repeat the following sequence until the stopping condition is met:
(l) Set k ← 1;
(2) *Repeat* the following steps until k =k_{max}.
(a) *Shaking.* Generate a point *x'* at random from the k^{ith} neighborhood $N_k(x)$ of x;
(b) *Local search by VNS,*
(b1) *Set* $\ell \leftarrow 1$;
(b2) *Repeat* the following steps until $\ell = \ell_{max}$;
- *Exploration of neighborhood.* Find the best neighbor *x''* of x' in $N\ell(x')$;
- *Move or not. If* $f(x'') < f(x')$ set $x' \leftarrow x''$ and $\ell \leftarrow 1$; otherwise set $\ell \leftarrow \ell + 1$;

(c) *Move or not.* If this local optimum is better than the incumbent, move there (x ← x"); and continue the search with N_1(k←1) otherwise, set k← k+1;

Algorithm 2: modified K-mean with VNS

Cuckoo Search Algorithm

Cuckoo birds attract attention of many scientists around the world because of their unique behaviour. They have many characteristics which differentiate them from other birds, but their main distinguishing feature is aggressive reproduction strategy. Some species such as the *Ani* and *Guira*

cuckoos lay their eggs in communal nests, though they may remove others' eggs to increase the hatching probability of their own eggs. Cuckoos engage brood parasitism. It is a type of parasitism in which a bird (brood parasite) lays and abandons its eggs in the nest of another species. There are three basic types of brood parasitism: intra-specific brood parasitism, cooperative breeding, and nest takeover. [13]

Some host birds do not behave friendly against intruders and engage in direct conflict with them. In such situation host bird will throw those alien eggs away. In other situations, more friendly hosts will simply abandon its nest and build a new nest elsewhere. [14]

Modified K-mean with Cuckoo Search Algorithm

In the real world, if a cuckoo's egg is very similar to a host's eggs, then this cuckoo's egg is less likely to be discovered, thus the fitness should be related to the difference in solutions depends on the step size value (which it mean the probability to finding the best solution). In the other hand the original K-mean algorithm find the best solution by find the closest items in each cluster only. Therefore, it is a good idea to do find a closest items in each cluster by using the fitness measure in cuckoo search algorithm (in this case the cluster means the nests) by increasing the degree of nearest between the items in the clusters, through we discard the weakness nests which have the lower number of items after that regenerate randomly a new nests instead of the discarded nests with re-computing the fitness of all clusters After nests sorted in a matrix by fitness depending on the best solutions. In this way, higher fitness solutions have slight advantage over solutions with lower fitness. This method keeps the selection pressure (the degree to which highly fit solutions are selected) towards better solutions and algorithm should achieve better results.

1. Initialize n particles/nests (which mean number of clusters).
2. **initial step size (used to find the fitness)**.
3. Repeat till stopping criteria met.
   a) Calculate fitness (Fi) of each particle.
   b) global best position is the best fit particle.
   c) move all the particles towards the global best position.
   d) for each particle if (fitness of current position (Fi) < fitness of personal best (Fj)) then replace solution j (personal best = current position).
   e) update personal best position for each particle.
   f) global best fitness value is retained.
4. Cluster centre is the global best position
5. end.

Algorithm 3: modified k-mean with cuckoo search.

Results and Discussion

In this section, shows experimental results which validate the modified Variable Neighbourhood Search (VNS) algorithm as a metahieuristic search and (CS) algorithm as a swarm intelligence search with k-mean clustering. As mentioned above, we developed our database clustering software, and all tests were run in the testing environment.

For testing purposes, an implementation original version of k-mean algorithm with variant clusters number was done to compare its results in term of the efficiency and performance with the proposed algorithms results.

To perform the testing perfectly, three databases form University of California Irvine (UCI) machine learning repository website [15] (Car, Wine and Zoo), in addition to two local databases (Human resources and IC3 exams) was used.

Car Evaluation Data Set

Abstract: Derived from simple hierarchical decision model, this database may be useful for testing constructive induction and structure discovery methods.

- Data Set Characteristics: Multivariate
- Number of Instances: 1728
- Area: N/A
- Attribute Characteristics: Categorical
- Number of Attributes: 6

- Date Donated: 1997-06-01
- Associated Tasks: Classification
- Missing Values? No
- Number of Web Hits: 219795

Data Set Information:

Car Evaluation Database was derived from a simple hierarchical decision model originally developed for the demonstration by DEX, by M. Bohanec, V. Rajkovic: On expert system for decision making. The model evaluates cars according to the following concept structure:

Attribute Information:

1) Class Values: unacc, acc, good, vgood.
2) buying: vhigh, high, med, low.
3) maint: vhigh, high, med, low.
4) doors: 2, 3, 4, 5more.
5) persons: 2, 4, more.
6) lug_boot: small, med, big.
7) safety: low, med, high.

Wine Data Set

Abstract: Using chemical to analysis determine the origin of wines

- Data Set Characteristics: Multivariate
- Number of Instances: 178
- Area: Physical
- Attribute Characteristics: Integer, Real
- Number of Attributes: 13
- Date Donated: 1991-07-01
- Associated Tasks: Classification
- Missing Values? No
- Number of Web Hits: 279485

Data Set Information:

These data are the results of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars. The analysis determines the quantities of 13 constituents found in each of the three types of wines. The attributes are:

1) Alcohol
2) Malic acid
3) Ash
4) Alcalinity of ash
5) Magnesium
6) Total phenols
7) Flavanoids
8) Nonflavanoid phenols
9) Proanthocyanins
10) Color intensity
11) Hue
12) OD280/OD315 of diluted wines
13) Proline

Zoo Data Set

- Data Set Characteristics: Multivariate
- Number of Instances: 101
- Area: life
- Attribute Characteristics: Categorical, Integer
- Number of Attributes: 17
- Date Donated: 1990-05-15
- Associated Tasks: Classification
- Missing Values? No
- Number of Web Hits: 72552

Data Set Information:

A simple database contains 17 Boolean-valued attributes. The "type" attribute appears to be the class attribute.

Attribute Information:

1. animal name: Unique for each instance
2. hair: Boolean
3. feathers: Boolean
4. eggs: Boolean
5. milk: Boolean
6. airborne: Boolean
7. aquatic: Boolean
8. predator: Boolean
9. toothed: Boolean
10. backbone: Boolean
11. breathes: Boolean
12. venomous: Boolean
13. fins: Boolean
14. legs: Numeric (set of values:{0,2,4,5,6,8})
15. tail: Boolean
16. domestic: Boolean
17. catsize: Boolean
18. type: Numeric (integer values in range [1,7]).

Will explain some results getting from applying the proposed algorithms clustering on databases in some details:

Table (1)
k-mean modified using VNS result with two clusters and squared error function.

cluster1	cluster2	SEF
1960	1979	745
1962	1980	700
1961	1979	680
1961	1979	680

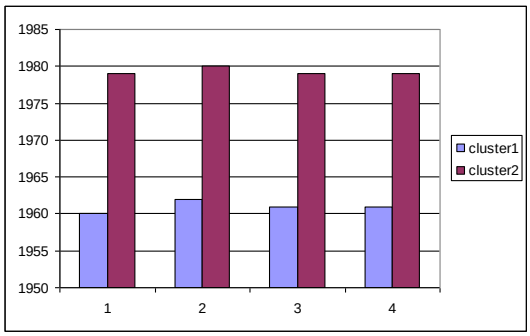

Fig. (1) k-mean modified using VNS result with two clusters.

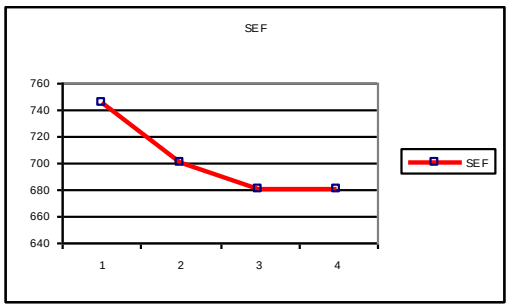

Fig. (2) Squared error function (SEF) with k-mean modified using VNS result with two clusters.

Table (2)
k-mean modified using VNS result with four clusters and squared error function.

Cluster1	Cluster2	Cluster3	Cluster4	SEF
1956	1966	1974	1983	737
1957	1968	1976	1983	725
1957	1967	1975	1984	700
1957	1969	1977	1984	685
1957	1969	1978	1984	675
1958	1969	1976	1984	660

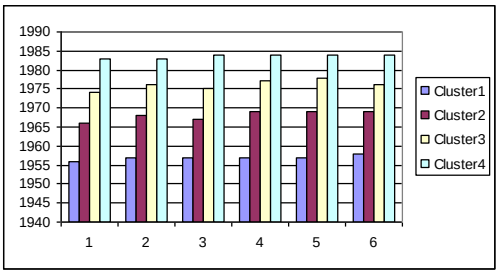

Fig.(3) k-mean modified using VNS result with four clusters.

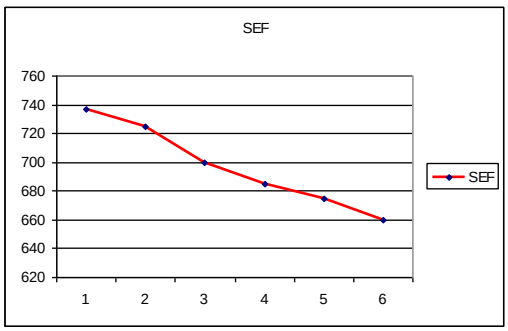

Fig.(4) Squared error function (SEF) with k-mean modified using VNS result with four clusters.

Table (3)
k-mean modified using CS result with four clusters and squared error function.

Cluster1	Cluster3	Cluster4	Cluster5	Cluster7	Cluster8	Cluster9	SEF
1956	1960	1966	1972	1974	1980	1983	760
1957	1961	1968	1973	1976	1981	1983	725
1957	1963	1967	1972	1975	1980	1984	700
1957	1962	1969	1973	1977	1979	1984	685
1957	1961	1969	1973	1978	1980	1984	675
1958	1961	1969	1973	1976	1980	1984	655
1958	1961	1969	1973	1976	1980	1984	640

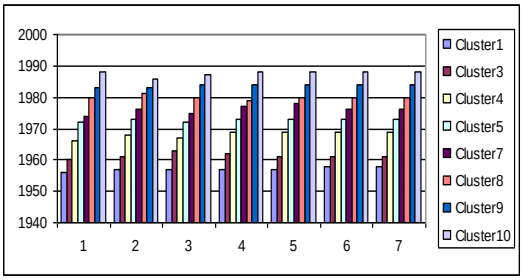

Fig.(5) k-mean modified using CS result with four clusters.

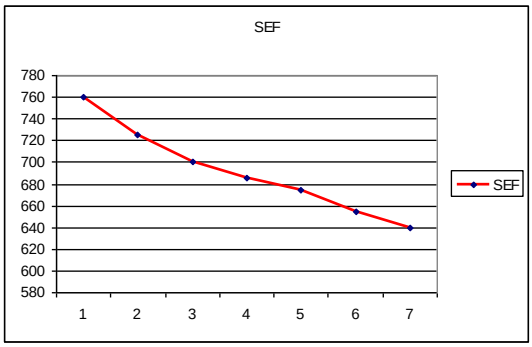

Fig.(6) Squared error function with k-mean modified using CS result with four clusters.

Conclusions

One of the challenges of clustering in data mining is the dealing with large databases. While many clustering algorithms work only on a small database, one solution is to use samples of the large database and cluster them, the quality of this clustering depends on samples, which may lead to biased results. Taking the whole large database will make the execution time too long to complete clustering. The suggested algorithms–that can deal with large database in a scalable manner– does not depend on sampling, thus it will give us real results (not biased). This algorithm generates the cluster point automatically by VNS and CS then applies on the k-mean for some iteration, which leads to minimize and stabilize the time of clustering.

From the experiment results stated in this paper, the following conclusions can be drawn:

1- The proposed algorithms do not need to input the value of number of clusters to it, instead the algorithm needs to pass automatically and randomly a new value in accurate way to determines the closeness between items in the same cluster. so also are the items in the same cluster are very close to each other.
2- Using metahieuristic methods for optimization solution such as VNS proposed algorithm, we could reduce the time needed to complete process from initial state to find the best solution (determine the best number of cluster) to half time or less.
3- The proposed algorithm (VNS) it affects the quality of clustering (performance) as it determines the closeness of items in the same cluster.
4- Using Swarm intelligence methods for optimization solution such as Cuckoo Search proposed algorithm, reduces the time which is needed to complete process from initial state to finding the best solution (determine the best number of cluster) to half time or less.
5- The proposed algorithm (CS) affects the quality of clustering (performance) since it determines the closeness of items in the same cluster.

References

[1] Jain A.K., Murty M.N. and Flynn P.J., "Data Clustering: A Review, ACM Computing Surveys", Vol. 31, No. 3, September 1999.

[2] Osama Abu Abbas, "Comparison between data clustering algorithms", the international arab journal of information technology, Vol. 5, No. 3, July 2008.

[3] Khaled Alsabti, Sanjay Ranka and Vineet Singh, "An Efficient K-Means Clustering Algorithm", Department of Computer and Information Science and Engineering, university of florida 1997.

[4] Subhash Sharma and Ajith Kumar, chapter 18 "cluster analysis and factor analysis" database management system, 2002.

[5] Johannes Grabmeier and Andreas Rudolph, "Techniques of Cluster Algorithms in Data Mining", 2002.

[6] Milan Tuba, "Swarm Intelligence Algorithms Parameter Tuning", Proceedings of the American Conference on Applied Mathematics, Cambridge, USA, 2012.

[7] Eberhart, R., Y. Shi, and J. Kennedy, "Swarm Intelligence", Morgan Kaufmann, San Francisco, 2001.

تقترح هذه الأطروحة طريقتين جديدتين لعنقدة قواعد البيانات، الأولى تقوم بتطوير خوارزمية العنقدة K-Mean Clustering من خلال دمجها مع خوارزمية البحث باستخدام التجاور او التقارب Variable Neighborhood Search، والثانية تقوم بتطوير خوارزمية العنقدة K-Mean Clustering من خلال دمجها مع خوارزمية بحث طائر الواق واق Cuckoo Search، حيث إن الخوارزميات المقترحة لا تحتاج إلى إدخال القيم يدويا إلى نقاط التمركز (نقاط العنقدة) حيث يتم إيجاد هذه القيم بشكل تلقائي للحصول على أفضل المجموعات مع استخدام خوارزميات التدقيق أو سماحية مجموعات العنقدة.

تم إجراء التجارب على العديد من قواعد البيانات ذات أحجام مختلفة، البعض منها تم الحصول عليها من الموقع الالكتروني لجامعة كاليفورنيا في ايرفين والخاصة بتعلم الآلة Machine Learning التي تحتوي على ٢٤٦ مستودع لقواعد البيانات والتي تقدم كخدمة للمجتمع لغرض استخدامها كأمثلة جاهزة لقواعد البيانات لتعلم الآلة. من هذه التجارب، تم استخلاص أن هذه الطرق المقترحة تقوم على تقليل الوقت الذي نحتاجه للحصول على أفضل حل في نصف الوقت الذي نحتاجه لتنفيذ الإجراءات نفسها باستخدام طرق العنقدة الأخرى. وفي الوقت نفسه استطعنا تخفيض التكرار للحصول على أفضل حل. وبالإضافة إلى ذلك، أن الطرق المقترحة تعطي أفضل جودة (كأداء) مقارنة مع طرق العنقدة الأخرى، حيث قمنا بتحسين الأداء بين (١٠٪ – ٢٠٪) مقارنة بأداء الخوارزمية الأصلية.

[8] Glover F., Gary Kochenberger A., "Handbook of Metaheuristics", Kluwer Academic, 2003.

[9] Mihai Gavrilas, "Heuristic and Metaheuristic Optimization Techniques with Application to Power Systems", Technical University of Iasi, D.Mangeron Blvd., Iasi, Romania, 2010.

[10] Abdulrahman Al-Guwaizani, "Variable Neighbourhood Search Based Heuristic for K-Harmonic Means Clustering", Department of Mathematical Sciences-School of Information Systems, Computing and Mathematics, Brunel University, London, May 2011.

[11] Chien-Yuan Tsai and Chuang-Cheng Chiu, "A VNS-based Hierarchical Clustering Method", Proceedings of the 5th WSEAS Int. Conf. on Computational Intelligence, Man-Machine Syatems and Cybernetics, Venice, Italy, 2006.

[12] Pierre Hansen and Nenad Mladenovic, "Variable Neighborhood Search", Chapter 8, Principles and applications, 2001.

[13] Yang, X. S., and Deb, S. Engineering, "Optimisation by Cuckoo Search", Int. J. of Mathematical Modelling and Numerical Optimisation, Vol. 1, No. 4, 2010.

[14] Senthilnath J., "Clustering using Levy Flight Cuckoo Search", Proceedings of 17th International Conference on Bio-Inspired Computing: Theories and Applications, Advances in Intelligent Systems and Computing Volume 202, 2013.

[15] UCI Machine Learning Database Repository, http://www.ics.uci.edu/~mlearn/MLSummary.html

الخلاصة

نتيجة للتطور الهائل في استخدام قواعد البيانات وامتلاك كميات كبيرة وضخمة من البيانات التي يتم جمعها في قواعد البيانات، فأن عملية التحليل العنقودي (العنقدة) لقواعد البيانات أصبحت في الآونة الأخيرة موضوعا مهما للغاية في مجال البحث عن واستخراج البيانات. ففي مجال التنقيب عن البيانات تركزت الجهود على إيجاد طرق أكثر كفاءة وفعالية في تحليل وعنقدة قواعد البيانات الكبيرة.