A Parameter-free Clustering Algorithm based K-means

Article *in* International Journal of Advanced Computer Science and Applications · January 2021

DOI: 10.14569/IJACSA.2021.0120372

Said Chah Slaoui
Faculté des sciences de Rabat.

Zineb Dafir
Mohammed V University of Rabat

A Parameter-free Clustering Algorithm based K-means

Said Slaoui[1], Zineb Dafir*[2]
Faculty of Science of Rabat, Mohammed V University
Rabat, Morocco

***Abstract*—Clustering is one of the relevant data mining tasks, which aims to process data sets in an effective way. This paper introduces a new clustering heuristic combining the E-transitive heuristic adapted to quantitative data and the k-means algorithm with the goal of ensuring the optimal number of clusters and the suitable initial cluster centres for k-means. The suggested heuristic, called PFK-means, is a parameter-free clustering algorithm since it does not require the prior initialization of the number of clusters. Thus, it generates progressively the initial cluster centres until the appropriate number of clusters is automatically detected. Moreover, this paper exposes a thorough comparison between the PFK-means heuristic, its diverse variants, the E-Transitive heuristic for clustering quantitative data and the traditional k-means in terms of the sum of squared errors and accuracy using different data sets. The experiments results reveal that, in general, the proposed heuristic and its variants provide the appropriate number of clusters for different real-world data sets and give good clusters quality related to the traditional k-means. Furthermore, the experiments conducted on synthetic data sets report the performance of this heuristic in terms of processing time.**

Keywords*—Data mining; clustering; overlapping clustering; k-means; cluster centre initialization

I. Introduction

In the last few years, the digital world has been facing rapid and unprecedented global evolutions due to the emergence of various concepts such as the development of the connected objects market, known as the internet of things, the continued growth of social networks, the strong use of the large e-commerce sites, as well as other factors. Therefore, this digital explosion presents a serious challenge for researchers to find appropriate techniques and efficient algorithms to analyze and process the considerable amount of data arising from those sources, and thus extract relevant information and facilitate decision-making.

Clustering is one of the relevant data mining tasks, which aims to process data sets effectively. Indeed, it proceeds by gathering the data objects the most similar into the same group, and the dissimilar ones into different groups, so that the similarity between data objects of the same group is the highest while the similarity between two data objects of different groups is the lowest [1]. The purpose is then to form disjoints groups, called clusters. The notion of similarity mainly depends on the attribute values describing the data objects and generally implies a distance measure. Accordingly, different clustering algorithms can make different clustering results for the same data set.

This paper suggests a parameter-free clustering algorithm combining the E-transitive heuristic [2] and the traditional k-means algorithm [3] [4]. Indeed, the proposed heuristic does not require the prior initialization of the number of clusters. It generates progressively the initial cluster centres until the appropriate number of clusters is automatically detected. Hence, the improvements achieved through this heuristic concern primarily two major weaknesses of the k-means algorithm. The first one consists of fixing the number of clusters k, and the second one focuses on the determination of the initial cluster centres. Moreover, an overall comparison was established between the parameter-free clustering algorithm based k-means, its diverse variants, the E-transitive heuristic [2] adapted to quantitative data, the iterative k-means minus-plus [5] and the traditional k-means [3] [4] in terms of the sum of squared errors and accuracy measures using different UCI data set [6]. The experiment results reveal that, in general, the proposed heuristic and its variants provide the appropriate number of clusters for different real-world data set and give good clusters quality compared to the traditional k-means. Furthermore, the experiments conducted on synthetic data sets report the performance of this heuristic in terms of processing time.

The major contributions of this paper are as follows:

- Develop a new parameter-free clustering heuristic combining the E-transitive heuristic and the k-means algorithm to automatically determine the initial cluster centres and the number of clusters.
- Provide different variants of the PFK-means algorithm focusing on the initialization process applied with different approaches (Overlapping PFK-means and Hard PFK-means).
- Establish a comparison between the suggested heuristic, its variants, the E-transitive adapted to quantitative data, the traditional k-means, and the iterative k-means minus-plus algorithm.

The remainder of this paper is organized as follows: Section 2 fully describes the different steps of the proposed heuristic, its different variants, and the E-transitive heuristic adapted to quantitative data. Section 4 provides the experiment results on real-world and synthetic data sets. And finally, Section 5 covers the conclusion and future perspectives.

II. Literature Review

Several clustering algorithms have been developed with the goal of ensuring optimal solutions for different clustering problems [7] [8] [9] [10] [11]. K-means is one of the popular partitioned clustering algorithms [3] [4], which aims to cluster a set of data objects into k clusters by minimizing the sum of squared errors over these clusters. Despite its popularity, efficiency, and facility of implementation, the major difficulty encountered with the k-means algorithm is primarily related to its sensitivity to the initialization conditions including the selection of the initial clustering centres, the determination of the number of clusters k, and the possibility to converge on a local optimum [12] [1]. All these aspects influence the quality of clustering. To deal with these issues, researchers devote continuously great efforts to find adequate techniques able to provide suitable initialization parameters, so then ensure a higher clustering quality. Diverse initialization improvements were suggested over the years such as [13] [12] [14][15] [16] [17] [5] [18] [19] [20].

Among these enhancements, the global k-means [13] which is considered as a global search procedure aiming to find an optimal solution for a clustering problem. Indeed, this proposed technique proceeds by adding dynamically one cluster centre at a time using a series of local searches based on fast computed bound on the clustering error. Moreover, it consists of splitting the data space using a k-d tree structure to improve the performance of clustering.

Another initialization strategy represented in cluster center initialization algorithm (CCIA) [12], which intends to perform clustering using two major steps. The first one consists of generating clusters whose number may exceed the number of clusters k. In such a case the second step is employed by merging the similar clusters using a density-based multi-scale data condensation, and then the merged clusters are treated as the initial cluster centres of the k-means algorithm.

In the same context, the k-means ++ algorithm [14] aims to select the initial cluster centre uniformly at random, then choose the next cluster centres based on a determined probability until the total number of clusters is reached. The next step consists of applying the standard k-means algorithm.

Further enhancement regarding the k-means initialization strategies manifested in the modified global k-means algorithm [15] intents to compute clusters incrementally and determine the k-partition of the data set used based on the previous iterations. Thereby, the algorithm calculates the starting points by minimizing an auxiliary cluster function.

In a similar vein, the authors in [17] develop a new canopy clustering; a pre-processing method for the k-means algorithm, which aims to determine appropriate initial clustering centers and thus attains an optimal number of clusters k. The proposed algorithm covers the pre-processing density canopy method as well as the main k-means processes.

More recently, an entropy-based initialization method [18] for the k-means algorithm was developed to obtain an optimal number of clusters. Indeed, it determines the initial point using the maximization of Shannon's entropy-based objective function, then it aims to detect the best number of clusters based on the optimal cluster detection algorithm for faster convergence.

Another recent work represents the random initialization method [21] merging the bootstrap technique with the data depth concept. Thereby, this method employs k-means with bootstrap replications to find the cluster centres in the original data space. Moreover, it aims to identify a good separation among clusters using depth computation.

III. A Parameter-free Clustering Algorithm based K-means

A. Basic Concepts

Suppose a data set$X$={ x_1, ..., x_n }, containing n data objects in Euclidean space, x_i∈R^d, i=1...n. The aim is to partition X into k clusters C_1, C_2, ...,C_k, that is, $\bigcup_{j=1}^{k} C_j$=X and $C_i \cap C_j$=$\emptyset$ for 1≤ i≠j ≤ k. c_1,c_2,...,c_k are the centres of clusters C_1, C_2, ...,C_k respectively and c_j=$\frac{1}{|C_j|}\sum_{x_i \in C_j} x_i$.

The difference between c_i, a centre of cluster C_j and a data object x is measured by $dist(x, c_i)$, where $dist(x, y)$ is the Euclidean distance between two data objects x and y. The quality of cluster C_j can be measured by the sum of squared error between all data objects x_i in C_j and the cluster centre c_j, defined as:

$$E = \sum_{i=1}^{n} \sum_{j=1}^{k} dist(x_i, c_j)^2, x_i \in C_j \qquad (1)$$

The average distance of all data objects in the data set$X$ is defined as follows:

$$MeanDist(X) = \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j=i+1}^{n} dist(x_i, x_j) \qquad (2)$$

A new cluster centre $c_{new} \in X$ corresponds to the data object defined by:

$$dist(c_{new}, c_j) = Max(dist(x_i, c_j)) \qquad (3)$$

where 1≤j≤p, p≤k, $x_i \in X$, p is the number of the existing cluster centres { C_1, ..., C_p } and c_j corresponds to the jth cluster centre.

B. The PFK-means Heuristic

The proposed heuristic is a parameter-free clustering algorithm, named PFK-means, combining the E-transitive heuristic [2] adapted to quantitative data and the traditional k-means [3][4]. Indeed, PFK-means does not require any initial parameters and generates progressively the cluster centres until the appropriate number of clusters is automatically detected. More specifically, the PFK-means consists of two major stages: the first stage includes the construction of the initial cluster centres and thus discovers the number of clusters k. The second stage consists of applying the traditional k-means algorithm by taking the cluster centres of the first stage as well as the number of clusters detected in the previous stage.

1) The initialization stage: This stage aims to establish the cluster centres without specifying the number of clusters k. In that respect, it starts by calculating the average distance of all data objects using the equation 2. Then, it selects the first cluster centre randomly from the data set containing n data objects. The next step consists of calculating the distance between the selected centre and each data object in the data set, using the Euclidean distance. In the case where the distance value is less than the average distance value, the corresponding data object is added to the overlapping cluster, which is being formed. Otherwise, no changes will be applied.

The selection of the other cluster centres follows another strategy. In such a case, the selection is decided during the construction of the forgoing overlapping cluster. The data object the least similar to the foregoing cluster centres is defined as the cluster centre of the current overlapping cluster. In other words, the new cluster centre corresponds to the data object defined in equation 3. From the second iteration, after the determination of the cluster centre, the construction of the other overlapping clusters is made similarly to the first step. This process continues until all data objects are processed and thus the initial clusters, as well as the overlapping clusters, are obtained. The steps above-mentioned are described in the algorithm 1.

Algorithm 1 Construction of initial clusters

```
Input:A set of $n$ data objects $X$
Output:The initial cluster centres $Tc_{next}$. The number of clusters automatically computed
begin
  compute $MeanDist(X)$ or $MeanDist(SampleX)$
  select the first cluster centre $c_{new}$ randomly from $X$
  initialize $Next \leftarrow$true
  $Tc_{next} \leftarrow$ null
  add $c_{new}$ to $Tc_{next}$
  while $Next$ do
      $c_{next} \leftarrow$ null
      for $i \leftarrow 0$ to $|X|$ do
          calculate $dist(x_i, c_{new})$
          if $dist(x_i, c_{new}) < MeanDist(X)$ then
              assign $x_i$ to the current cluster
          else if $c_{next}$ is null then
              $c_{next} \leftarrow x_i$
          else if $dist(x_i, c_{new}) > dist(c_{next}, c_{new})$ for all $c_{next}$ in $Tc_{next}$ then
              $c_{next} \leftarrow x_i$
          end if
      end for
      $c_{new} \leftarrow c_{next}$
      add $c_{new}$ to $Tc_{next}$
      if $c_{next}$ is null then
          $Next \leftarrow$ false
      end if
  end while
end begin
```

2) *The second stage:* The purpose of the initialization stage is to provide the initial cluster centres and detect the number of clusters k automatically. These parameters are the input settings of the traditional k-means executed in this stage. In that respect, the procedure starts by browsing the whole data set and thereafter scrolls through the list of cluster centres provided from the initialization stage and finally assign each data object to the appropriate cluster according to the Euclidean distance. After assigning all data objects to the appropriate clusters, the cluster centres are updated by calculating the mean of the data objects contained in each cluster. The process reiterates until there is no change in the cluster centres values. It should be noted that using the initial cluster centres obtained in the initialization stage as input settings of the traditional k-means allows a rapid convergence and an optimum solution. The pseudo-code of the traditional k-means is described in Algorithm 2.

Algorithm 2 The traditional k-means

```
Input: a set of $n$ data objects $X$, the list of initial clusters $Tc_{next}$, the number $k$ automatically computed
Output: the data objects in $X$ partitioned in $k$ clusters
begin
  repeat
      for $i \leftarrow 0$ to $|X|$ do
          for $j \leftarrow 0$ to $k$ do
              calculate $dist(x_i, c_j)$
              if $dist(x_i, c_j) < MeanDist(X)$ then
                  assign $x_i$ to the current cluster
              end if
          end for
      end for
      update the cluster centres
  until Convergence criteria are met
end begin
```

C. *Different Variants of PFK-means*

In order to fully explore the suggested heuristic, several variants of PFK-means have been proposed. These variants mainly focus on the initialization process applied with different approaches: overlapping PFK-means and hard PFK-means.

1) *Overlapping PFK-means variant:* In the initialization stage, each data object can belong to several clusters and thus the obtained distribution contains overlapping clusters. In that respect, there is one suggested solution with overlapping clusters.

In order to obtain the initial clusters and the number of clusters, the first variant of PFK-means consists of applying the initialization procedure, which is above-explained as a first stage. Thereafter, the second stage starts by browsing the achieved cluster centres and the whole data set and for each data object, calculates the distance between this data object and the current cluster centre based on the Euclidean distance. In the case where the distance value is less than the average distance of all data objects (equation 2), the data object being processed is added to the overlapping cluster which is being formed. After scanning the whole cluster centres, each cluster centre value is updated by calculating the mean value of all data objects belonging to its corresponding cluster. This iterative procedure is repeated until no changes occur on the cluster centres values. After the completion of this process, the data set processed is partitioned into k overlapping clusters. The algorithm 3 shows the details of this iterative procedure.

2) *The hard PFK-means version I:* This solution consists of applying the initialization process ((Algorithm 1) presented in the PFK-means heuristic as a first stage. Then, similarly to the steps explained on the iterative procedure (Algorithm 3) assign each data object to the appropriate cluster and in parallel remove the intersections between the constructed overlapping clusters. In other words, when the data object which is being processed is not clustered, it is added immediately to the cluster being formed. Otherwise, when the data object is already clustered, the distance between the current cluster centre and the data object is calculated then compared with

Algorithm 3 The iterative procedure

```
Input: a set of $n$ data objects $X$, the $k$ initial cluster centres $C$, $k$
Output: the data objects in $X$ partitioned in $k$ clusters
begin
  repeat
      for $r \leftarrow 0$ to $k$ do
          for $i \leftarrow 0$ to $|X|$ do
              calculate $dist(x_i, c_r)$
              if $dist(x_i, c_r) < MeanDist(X)$ then
                  assign $x_i$ to the current cluster
              end if
          end for
      end for
      update the cluster centres $C$
  until Convergence criteria are met
end begin
```

the distance between the same data object and the centre of the cluster containing this data object. In the case where the data object is most similar to the current cluster, it will be removed from the old cluster and added to the overlapping cluster being formed. Thus, the cluster centres are updated after each iteration by calculating the mean value of the data objects assigned to each cluster. The Algorithm 4 illustrates the steps of this hard iterative procedure.

Algorithm 4 The hard iterative procedure

```
Input: a set of $n$ data objects $X$, the $k$ initial cluster centres $C$, $k$
Output: the data objects in $X$ partitioned in $k$ clusters
begin
  repeat
      for $r \leftarrow 0$ to $k$ do
          for $i \leftarrow 0$ to $|X|$ do
              if $x_i$ is not clustered then
                  calculate $dist(x_i, c_r)$
                  if $dist(x_i, c_r) < MeanDist(X)$ then
                      assign $x_i$ to the current cluster
                  end if
              else if $x_i$ is clustered in the cluster whose centre is $c_m$ then
                  calculate $dist(x_i, c_r)$ and $dist(x_i, c_m)$
                  if $dist(x_i, c_r) < dist(x_i, c_m)$ then
                      add $x_i$ to the current cluster, remove $x_i$ from the cluster represented by $c_m$
                  end if
              end if
          end for
      end for
      update the cluster centres $C$
  until Convergence criteria are met
end begin
```

3) The hard PFK-means version II : The process of this variant is similar to that of the first version of the hard PFK-means (Algorithm 1+ Algorithm 4), the only difference is that the last stage of this solution consists of applying the traditional k-means at the end of the process. In that regard, the second version of the hard PFK-means consists of three major stages. The first stage aims to discover the initial clusters by applying the initialization phase as presented in the PFK-means (Algorithm 1). Then, the second stage consists of executing the hard iterative procedure as explained in the above variant (Algorithm 4). Finally, in the last stage, the traditional k-means is applied by taking the cluster centres and the number of clusters, obtained from the second stage, as input parameters (Algorithm 2).

4) The hard PFK-means version III: In a similar vein, this solution starts by applying the initialization stage (Algorithm 1). Secondly, it executes the iterative procedure (Algorithm 3). At last stage, it runs the traditional k-means algorithm taking as input settings the output parameters of the second stage, which are the initial cluster centres of the obtained overlapping clusters and the number of overlapping clusters automatically computed (Algorithm 2).

D. The E-transitive Heuristic Adapted to Quantitative Data

The E-transitive heuristic [2] is an improved version of the *Transitive* heuristic [22] which aims to cluster categorical data sets using the benefits of the Relational Analysis [23]. In fact, the principal purpose of this heuristic is to perform a clustering without specifying the number of clusters by adopting a specific cluster structure and then reduce the computational time. Thus, the E-transitive heuristic adapted to quantitative data consists of applying exclusively the initialization stage (Algorithm 1) presented in the PFK-means heuristic by removing intersections between the overlapping clusters. In that regard, the process is similar to that of the initialization stage, the only difference is that each data object must be checked before being added to the appropriate cluster. Thus, at the beginning of the process, all data objects are noted as not clustered and each data object added to a cluster is noted clustered. Accordingly, there are two possibilities. In the case where the data object being processed is not clustered, it will be added to the cluster being formed immediately. Otherwise, in the first step, the distance between the current cluster centre and the data object is calculated and then compared with the distance between the same data object and the centre of the cluster containing this data object. In the case where the data object is most similar to the current cluster, the data object will be removed from the old cluster and added to the overlapping cluster being formed. The cluster centres are updated after each modification. The Algorithm 5 presents the instructions of this solution.

IV. Experiments

This section provides the results obtained by implementing the PFK-means heuristic, its different variants, the E-transitive heuristic [2] adapted to quantitative data, and the traditional k-means [3] [4] using real-world data sets, retrieved from the UCI Machine Learning Repository [6]. In order to measure the clustering effect, these algorithms are evaluated based on the accuracy and the sum of squared errors described by equation 1. The experiments include also the simulation tests which have been performed to evaluate the performance of PFK-means heuristic in terms of running time with distinct synthetic data sets generated using a data mining generator, called weka [24].

Algorithm 5 The E-transitive heuristic adapted to quantitative data

```
Input:A set of $n$ data objects $X$
Output:The initial cluster centres $Tc_{next}$. The number of clusters automatically computed
begin
compute $MeanDist(X)$ or $MeanDist(SampleX)$
select the first cluster centre $c_{new}$ randomly from $X$
initialize $Next$ ←true
$Tc_{next} \leftarrow$ null
add $c_{new}$ to $Tc_{next}$
while $Next$ do
    $c_{next} \leftarrow$ null
    for $i \leftarrow 0$ to $|X|$ do
        calculate $dist(x_i, c_{new})$
        if $dist(x_i, c_{new}) < MeanDist(X)$ then
            if $x_i$ is not clustered then
                assign $x_i$ to the current cluster
                update the current cluster
            else if $x_i$ is clustered in the cluster whose centre is $c_m$ then
                calculate $dist(x_i, c_r) anddist(x_i, c_m)$
                if $dist(x_i, c_r) < dist(x_i, c_m)$ then
                    add $x_i$ to the current cluster, remove $x_i$ from the cluster represented by $c_m$
                    update the current cluster and cluster represented by $c_m$
                end if
            end if
        else if $c_{next}$ is null then
            $c_{next} \leftarrow x_i$
        else if $dist(x_i, c_{new}) > dist(c_{next}, c_{new})$ for all $c_{next}$ in $Tc_{next}$ then
            $c_{next} \leftarrow x_i$
        end if
    end for
    $c_{new} \leftarrow c_{next}$
    add $c_{new}$ to $Tc_{next}$
    if $c_{next}$ is null then
        $Next \leftarrow$ false
    end if
end while
end begin
```

A. Data Sets Description

Table I gives a brave description of seven real-world data sets, retrieved from the UCI machine learning [6], used to evaluate the performance of the proposed heuristic, namely, Iris, Wine, Seeds, Pima Indian Diabetes, Soybean-small, Segmentation, Musk, and Letter-Recognition (LR). As shown in Table I, each data set is described by a specified number of clusters, many data objects, and each data object is described by a vector of attributes. The simulated data sets are generated by varying the size of the data sets, the number of clusters, and the number of attributes. Indeed, the first experiment consists of generating data sets with different sizes: 1000, 1500, 2000, 2500, 3000, 3500, and 4000. Each of these data sets is described by three clusters and five attributes. In the second experiment, the data set size is fixed at 1000, the number of attributes at 5, and the size of the cluster is varied as follows: 2, 3, 4, 5, 6, 7, 8, 9, and 10. Finally, the last experiment use data sets with a different number of attributes: 5, 10, 15, 20, 25, 30, 35, 40, and fix the data set size and the number of clusters at 1000 and three respectively. Besides, two-dimensional synthetic data sets [25] were used for comparing the suggested heuristic and the iterative k-means minus-plus [5]. These data sets contain 5000 data poins and 15 clusters: S1, S2, S3, and S4.

TABLE I. DESCRIPTION OF THE REAL-LIFE DATA SETS USED IN THE EXPERIMENTS.

data set	Data size	Attributes	Cluster number
Iris	150	4	3
Wine	178	13	3
Soybean-small	47	35	4
Pima Indian Diabetes	768	8	2
Seeds	210	7	3
Musk	6598	168	2
Letter-Recognition(LR)	20000	16	26

B. Clustering Evaluation Measures

Clustering validation is an important aspect to evaluate the quality of clustering results. Indeed, it depends on some parameters such as the similarity measure, the implementation of the clustering algorithm used, and the capacity to catch some or all of the hidden patterns. In order to measure the clustering effect of the proposed heuristic, the following parameters are involved: the time required for completing the procedure of clustering, the sum of squared errors (equation 1), the accuracy, and the entropy clustering measure.

C. Results on Real-world Data Sets

In order to evaluate the performance of the developed heuristic and its variants, these heuristics have been programmed using java. The results presented are the best values obtained from five runs for each proposed heuristic, except for the second version of the hard PFK-means which produces stable results. Concerning the traditional k-means algorithm, the initial centres were generated randomly. Table II provides the sum of squared errors for PFK-means, the E-transitive heuristic [2] adapted to quantitative data and its variants on real-world data sets. Clearly, the second version of the hard PFK-means exceeds the other proposed heuristics in terms of the sum of squared errors for all data sets except the soybean data set. In this case, the PFK-means and the second version of the hard PFK-means give the best results. The performance of the E-transitive heuristic comes back to the fact that this variant makes it possible to detect outliers. Absolutely, since in the iterative procedure applied as the second stage of this heuristic, the data objects which are very far from the cluster centres are not assigned imperatively to these clusters. Therefore, the E-transitive heuristic adapted to quantitative data provides the minimal values of SSE. Additionally, the second version of the hard PFK-means produces stable results. Finally, it should be noted that PFK-means and its variants lead to finding the right number of clusters for all tested data sets as described in Table IV. Furthermore, regarding the PFK-means heuristic and its hard variants, all inputs are clustered.

In the figures (Fig. 1, Fig. 2, Fig. 3), a comparison of the clustering results of PFK-means, the E-transitive heuristic

TABLE II. THE SUM OF SQUARED ERRORS OF THE CLUSTERING RESULTS ON REAL-WORLD DATA SETS.

data set	PFK-means	E-transitive	Hard PFK-means1	Hard PFK-means2	Hard PFK-means3
Iris	78.94	**76.46**	82.86	78.94	78.94
Soybean	**205.96**	207.49	220.05	207.49	207.05
Wine	2.63E+06	**2.31E+06**	2.97E+06	2.37E+06	2.37E+06
Pima	5.18E+06	**4.31E+06**	5.68E+06	5.12E+06	5.14E+06
Seeds	**587.31**	**587.31**	630.78	**587.31**	588.43

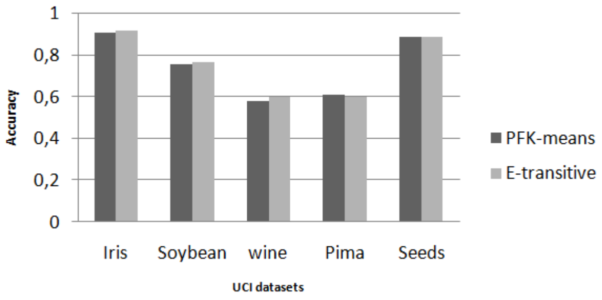

Fig. 1. The Accuracy of PFK-means and E-transitive on Real-world Data Sets.

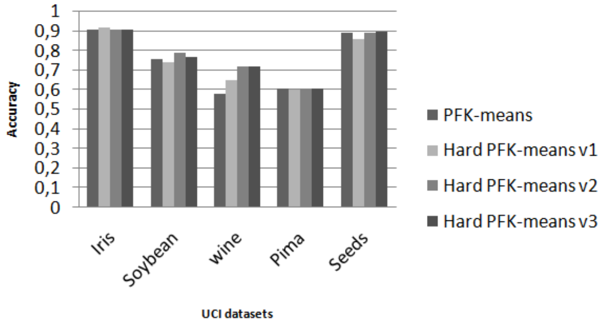

Fig. 2. The Accuracy of PFK-means and its Variants on Real-world Data Sets.

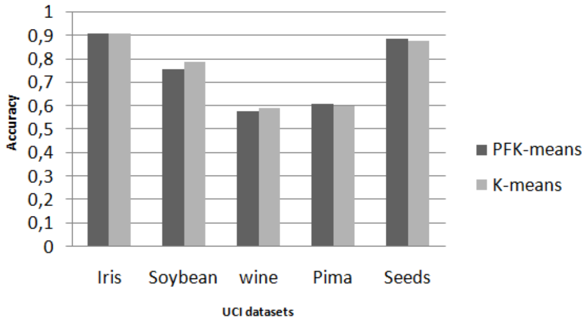

Fig. 3. The Accuracy of PFK-means and the Traditional k-means on Real-world Data Sets.

adapted to quantitative data, its variants, and the traditional K-means on real-world data sets in term of accuracy is illustrated. Fig. 1 presents a comparison between the PFK-means and the E-transitive heuristic. Based on this result, it is clear that the accuracy of these heuristics is closed to each other. Moreover, Fig. 2 describes the accuracy of PFK-means and its variants for the above-mentioned real-world data sets. As can be seen, the accuracy of PFK-means and its variants are closed to each other for iris, Pima Indian Diabetes, seeds, and soybean data sets, yet for wine data set the second version of the hard PFK-means and the third version of the hard PFK-means outperform the PFK-means heuristic and the first version of the hard PFK-means. The last Fig. (3) describes the accuracy of the PFK-means heuristic and the k-means algorithm with UCI data sets [6]. From this figure, it can be shown that the PFK-means outperforms the k-means algorithm for Pima Indian Diabetes and seeds data set. However, for wine and soybean data sets, the k-means algorithm achieves an accuracy superior to the accuracy of PFK-means.

Table III presents a comparison between the PFK-means heuristic and the traditional k-means Algorithm [3] [4], in terms of a sum of squared errors, accuracy, and entropy measure based on real-world data sets. Regarding the sum of squared errors and the entropy clustering measure, the smaller their values, the better the result. The highest value provided by the entropy clustering measure is one while the lowest one is 0. The PFK-means heuristic exceeds the traditional k-means algorithm in terms of the sum of squared errors for all data sets except the Pima data set. Concerning the accuracy, the PFK-means heuristic gives the best results for Iris, Pima, and Seeds data sets. Furthermore, the values obtained by the entropy clustering measure are the best for the PFK-means heuristic. In addition to that since the suggested heuristic doesn't require the number of clusters as an input parameter, it shows important results compared to the traditional k-means. Thus, the results exposed in Table III demonstrate the efficiency of the PFK-means heuristic and its ability to find the appropriate number of clusters for all data sets. Table IV presents the number of clusters found after executing the PFK-means as well as the exact number of clusters for the real-world data sets used.

TABLE III. COMPARISON OF PFK-MEANS AND K-MEANS ON REAL-WORLD DATA SETS.

	K-means			PFK-means		
	SSE	**Accuracy**	**Entropy**	**SSE**	**Accuracy**	**Entropy**
Iris	78.94	0.91	0.39	78.94	0.91	0.39
Soybean	208.15	**0.79**	0.56	**205.96**	0.76	**0.55**
Wine	2.66E+06	**0.59**	0.98	**2.63E+06**	0.58	**0.95**
Pima	**5.13E+06**	0.60	0.91	5.18E+06	**0.61**	**0.77**
Seeds	593.50	0.88	0.46	**587.31**	**0.89**	**0.44**

TABLE IV. THE NUMBER OF CLUSTERS OBTAINED AFTER EXECUTING PFK-MEANS AND ITS VARIANTS.

data set	Exact Cluster number	Cluster number found
Iris	3	3
Wine	3	3
Soybean-small	4	4
Pima Indian Diabetes	2	2
Seeds	3	3
Musk	2	2
Letter-Recognition(LR)	26	26

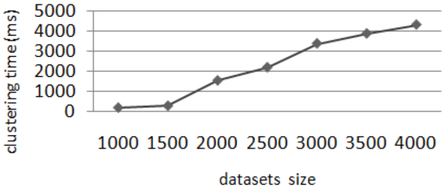

Fig. 4. Clustering Time of Synthetic Data Sets for Different Sizes.

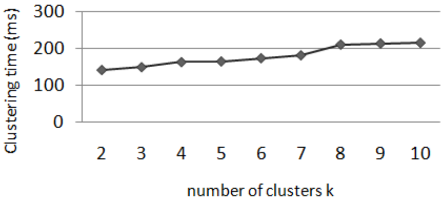

Fig. 5. Clustering Time of Synthetic Data Sets for Different Clusters.

D. Results on Synthetic Data Sets

For the purpose of testing the performance of the proposed heuristic in terms of processing time, different synthetic data sets were generated based on different clustering criteria namely, the size of the data sets, the number of clusters, and the number of attributes.

The first experiment (Fig. 4) describes the performance of the proposed heuristic when increasing the size of the data sets, which varies from 1000 to 4000 while setting the number of clusters at 3 and the number of attributes at 5. As shown in Fig. 4 the running times of the proposed heuristic vary from 188 to 4321 milliseconds which are nearly linear against the size of the data sets. The next experiment (Fig. 5) depicts the suggested heuristic behavior by increasing the number of clusters from 2 to 10 with the data set size set to 1000 instances and the number of attributes fixing to 5. It is clear from Fig. 5 that the clustering time scales linearly from 141 to 217 milliseconds while increasing the number of clusters. Additionally, the proposed heuristic can detect the adequate number of clusters of each generated data set. The last experiment (Fig. 6) illustrates the processing times in milliseconds while increasing the number of attributes from 5 to 40 with the number of clusters fixing to 3 and the size of the data set setting in 1000. This Fig. (6) shows clearly that the variation of the running times of the proposed heuristic from 169 to 489 milliseconds while increasing the number of attributes is quite linear.

E. Results of PFK-means Compared to the Iterative k-means Minus-plus

The PFK-means heuristic was compared to the iterative k-means minus-plus [5], which is an iterative approach to improve the quality of the k-means algorithm by removing one cluster (minus), dividing another one (plus), and applying re-clustering again, for each iteration. The results of the iterative k-means minus-plus and the traditional k-means were presented as in the original paper describing the iterative k-means [5]. Table V presents a comparison between the PFK-means heuristic, the iterative k-means minus-plus, and the traditional k-means algorithm, in terms of a sum of squared errors, and maximum of partial SSE for three real-world data sets and four synthetic data sets. The PFK-means heuristic and the iterative k-means minus-plus algorithm outperforms the traditional k-means, except for S3 and S4 data sets when the iterative k-means -+ outperforms the proposed heuristic.

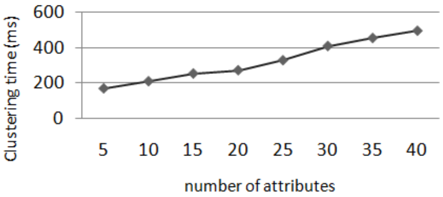

Fig. 6. Clustering Time of Synthetic Data Sets by Varying the Number of Attributes.

TABLE V. Comparison of PFK-means, the Iterative k-means-+, and the K-means Algorithm on Different Data Sets.

	maximum of partial SSE			SSE		
	KM	IKM-+	PFKmeans	KM	IKM-+	PFKmeans
Iris	6.53E+01	3.98E+01	3.98E+01	9.95E+01	7.89E+01	7.89E+01
Musk	4.61E+09	4.35E+09	4.35E+09	6.09E+09	5.92E+09	5.92E+09
LR	4.17E+04	3.93E+04	3.93E+04	6.20E+05	6.16E+05	6.16E+05
S1	6.59E+12	8.54E+11	8.54E+11	1.85E+13	8.92E+12	8.91E+12
S2	5.78E+12	1.33E+12	1.33E+12	2.01E+13	1.33E+13	1.32E+13
S3	3.41E+12	1.56E+12	2.97E+12	1.94E+13	1.69E+13	2.06E+13
S4	2.56E+12	1.79E+12	2.77E+12	1.70E+13	1.57E+13	1.98E+13

V. Conclusions and Perspectives

The purpose of this research is to present a new clustering algorithm namely a parameter-free clustering algorithm based on k-means. This hybrid solution combines the E-transitive heuristic adapted to quantitative data and the k-means algorithm to deal with the major issue encountered with k-means, which is the determination of the number of clusters and the initial cluster centres. The PFK-means and its variants were explained according to the clustering approaches. Also, this paper covers a detailed comparison between the PFK-means heuristic, its different variants, the revisited version of the E-transitive heuristic, the iterative k-means minus-plus, and the k-means algorithm in terms of the sum of squared errors and accuracy.

From the experiments that have been conducted on real-world data sets, it has been proven that the suggested heuristics can to detect the appropriate number of clusters independently of any initial conditions. Accordingly, these heuristics can be successfully used for unsupervised learning. Furthermore, the examination conducted on synthetic data sets demonstrates that the proposed heuristic finds the appropriate number of clusters in reasonable processing time against the variation of the size of the data sets, the number of clusters, and the number of attributes. In future work, we will be concentrating on clustering big data using the parallel programming [26] to improve the efficiency and the complexity of the proposed heuristic. Additionally, we will focus on the implementation of the proposed heuristic using other similarity measures.

REFERENCES

[1] J. Han, J. Pei, and M. Kamber, *Data mining: concepts and techniques*. Elsevier, 2011.

[2] S. C. Slaoui, Z. Dafir, and Y. Lamari, "E-transitive: an enhanced version of the transitive heuristic for clustering categorical data," *Procedia Computer Science*, vol. 127, pp. 26–34, 2018.

[3] J. MacQueen *et al.*, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, no. 14. Oakland, CA, USA, 1967, pp. 281–297.

[4] S. Lloyd, "Least squares quantization in pcm," *IEEE transactions on information theory*, vol. 28, no. 2, pp. 129–137, 1982.

[5] H. Ismkhan, "Ik-means-+: An iterative clustering algorithm based on an enhanced version of the k-means," *Pattern Recognition*, vol. 79, pp. 402–413, 2018.

[6] A. Asuncion and D. Newman, "Uci machine learning repository," 2007.

[7] P. Berkhin, "A survey of clustering data mining techniques," in *Grouping multidimensional data*. Springer, 2006, pp. 25–71.

[8] L. Rokach, "A survey of clustering algorithms," in *Data mining and knowledge discovery handbook*. Springer, 2009, pp. 269–298.

[9] A. K. Jain, "Data clustering: 50 years beyond k-means," *Pattern recognition letters*, vol. 31, no. 8, pp. 651–666, 2010.

[10] H. Zhang, T. W. Chow, and Q. J. Wu, "Organizing books and authors by multilayer som," *IEEE transactions on neural networks and learning systems*, vol. 27, no. 12, pp. 2537–2550, 2015.

[11] H. Zhang, S. Wang, X. Xu, T. W. Chow, and Q. J. Wu, "Tree2vector: learning a vectorial representation for tree-structured data," *IEEE transactions on neural networks and learning systems*, no. 99, pp. 1–15, 2018.

[12] S. S. Khan and A. Ahmad, "Cluster center initialization algorithm for k-means clustering," *Pattern recognition letters*, vol. 25, no. 11, pp. 1293–1302, 2004.

[13] A. Likas, N. Vlassis, and J. J. Verbeek, "The global k-means clustering algorithm," *Pattern recognition*, vol. 36, no. 2, pp. 451–461, 2003.

[14] D. Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding," in *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*. Society for Industrial and Applied Mathematics, 2007, pp. 1027–1035.

[15] A. M. Bagirov, "Modified global k-means algorithm for minimum sum-of-squares clustering problems," *Pattern Recognition*, vol. 41, no. 10, pp. 3192–3199, 2008.

[16] M. E. Celebi, H. A. Kingravi, and P. A. Vela, "A comparative study of efficient initialization methods for the k-means clustering algorithm," *Expert systems with applications*, vol. 40, no. 1, pp. 200–210, 2013.

[17] G. Zhang, C. Zhang, and H. Zhang, "Improved k-means algorithm based on density canopy," *Knowledge-based systems*, vol. 145, pp. 289–297, 2018.

[18] K. Chowdhury, D. Chaudhuri, and A. K. Pal, "An entropy-based initialization method of k-means clustering on the optimal number of clusters," *Neural Computing and Applications*, pp. 1–18, 2020.

[19] S. Xia, D. Peng, D. Meng, C. Zhang, G. Wang, E. Giem, W. Wei, and Z. Chen, "A fast adaptive k-means with no bounds," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[20] S. Wang, X. Liu, and L. Xiang, "An improved initialisation method for k-means algorithm optimised by tissue-like p system," *International Journal of Parallel, Emergent and Distributed Systems*, vol. 36, no. 1, pp. 3–10, 2021.

[21] A. Torrente and J. Romo, "Initializing k-means clustering by bootstrap and data depth," *Journal of Classification*, pp. 1–25, 2020.

[22] S. C. Slaoui and Y. Lamari, "Clustering of large data based on the relational analysis," in *2015 Intelligent Systems and Computer Vision (ISCV)*. IEEE, 2015, pp. 1–7.

[23] J. Ah-Pine and J.-F. Marcotorchino, "Overview of the relational analysis approach in data-mining and multi-criteria decision making," in *Web intelligence and intelligent agents*. IntechOpen, 2010.

[24] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[25] P. Fränti and S. Sieranoja, "K-means properties on six clustering benchmark datasets (2018)," *URL: http://cs. uef. fi/sipu/datasets*.

[26] Z. Dafir, Y. Lamari, and S. C. Slaoui, "A survey on parallel clustering algorithms for big data," *Artificial Intelligence Review*, pp. 1–33, 2020.