

Fuzzy Clustering of Large-Scale Data Sets Using Principal Component Analysis

Olfa Arfaoui

National Engineering School of Tunis
BP. 37, Le Belvédère 1002 Tunis, Tunisia
olfa.arfaoui@yahoo.fr

Minyar Sassi

National Engineering School of Tunis
BP. 37, Le Belvédère 1002 Tunis, Tunisia
minyar.sassi@enit.rnu.tn

Abstract—To effectively exploit large-scale data sets using a limited storage space, it is necessary to find a special treatment which reduces them. There are certain methods with this intention. We can quote clustering method. However, this method proves its limits in the case of large-scale data sets. In this paper, we propose to reduce the workspace using the Principal Component Analysis (PCA). We work with fuzzy clustering of a data set in which users don't know the optimal number of clusters to be generated. We proved the effectiveness of the pre-processing use of this technique before any clustering operation.

Keywords—Clustering; Fuzzy Logic; Principal Component Analysis; Sampling; Data Compression.

I. INTRODUCTION

The evolution of the data processing and storage technologies and tools involved an explosion of available data. This phenomenon was reinforced by the awakening of the decision makers of the potential knowledge value which contain their databases (DB). Nevertheless, the available large data don't bring any appreciation in absence of exploration and analysis methods. In this case, the installation of treatments tools to clarify the available knowledge in these data is fundamental and represents one of the major stakes of the Knowledge Discovery from Database (KDD).

The reduction of the space of description and the number of data are two dual approaches which bring brief replies to this problem. We often gather them under name "reduction of dimensions". Their objective is to select or extract an optimal subset from characteristics and/or objects for a fixed criterion. By eliminating non relevant or redundant information according to the criterion used, these approaches make it possible to reduce the size of the data set according to its two dimensions to make it more representative to a given problem.

They have the following objectives: (1) simplify the data visualization and comprehension; (2) reduce the storage space; (3) reduce the time of training and use, and (4) identify the relevant factors.

Several works were proposed in the literature to seek and improve the Data Mining methods among which we quote the Information Retrieval (IR), the data querying, the clustering, the Formal Concept Analysis (FCA) etc.

In this work, we focus ourselves on fuzzy clustering of large-scale data sets.

Several works improve this method while proving its performances in the management of the large-scale data sets that on the determination of the optimal number of clusters.

However, these solutions are insufficient to ensure a good clustering of the large-scale datasets, in which the data can have hundred attributes or variables and make the clustering in this space extremely difficult.

The Principal Component Analysis (PCA) permits to solve these problems while putting in considerations the following objectives: (1) represent observations moreover than 2 variables into 2 or 3 dimensions; (2) reduce dimension in a relevant way while guaranteeing a minimal loss of information.

The new variables obtained have the following characteristics: (1) they are ordered according to the information which it restores of the initial variables; (2) they are not linearly correlated between them.

The rest of the paper is organized as follows: section 2 presents some works on the clustering large data sets. Section 3 presents our approach for fuzzy clustering of large-scale data sets while using the PCA. Section 4 presents experimentation and evaluates the approach. Section 5 concludes the paper and presents the future work.

II. CLUSTERING OF LARGE DATA SETS

When we treat large datasets, the problem of the resources always emerges. In our case, the traditional clustering techniques have important complexities, so, it is not obvious to apply them to large datasets, because they are so enormous that we cannot store it in memory. A solution consists of applying compression technique to prevent that these kind of datasets exceeds the memory space.

For this, several techniques and approaches were applied to the methods and existing algorithms to deal with the problem of the clustering in the large datasets, we divided them into data reduction and dimension reduction techniques.

For data reduction, several researchers studied the concept of sampling and data compression [9], [10], [11], [12], [13], [14], [15], [16]. Table 1 presents a comparative study of these approaches.

TABLE I. COMPARATIVE STUDY OF LARGE DATA SETS CLUSTERING ALGORITHMS

Name	Type	Inputs	Used techniques (sampling or compression)	Optimized for
CURE	Hierarchical	$X+C$ + Number of representatives	Sampling	Large Data sets
CLARANS	Partition-based clustering	$X+C$ + Maximum number of the nearest	Sampling	Spatial data sets
GEFFCM	Partition-based clustering	$X+C$	Sampling	Large Data sets
Single pass k Means	Partition-based clustering	$X+C+V_i$	Compression	Large Data sets
Single pass fuzzy C-Means	Partition-based clustering	$X+C+P$	Compression	Large Data sets
WEFCM	Partition-based clustering	X	Compression	Large Data sets

Where X is the data set, C is the number of clusters, V the initial centers and P the sampling data.

The majority of the algorithms presented above are interesting, but they present some limits. Although they solve the problem of the clustering of large datasets, they require also the number of clusters like input parameter. However WEFCM (Weighted Enhance Fuzzy C-Means) [19] presents a good automatic system of clustering while provides the optimal number of clusters. Based on this point, and since this algorithm allows only the data reduction, we will associate to him a dimension reduction procedure. This is to facilitate the clustering of large-scale datasets.

The principle of this algorithm is as follow: suppose we intend to cluster a large data sets present on a disk. We can only load a certain percentage of the data based on the available memory allocation. If we load 1% of the data into memory at a time then we have to do it 100 times to scan through the entire data set. We call each such data access a Partial Data Access (PDA). The number of partial data accesses will depend on how much data we load each time.

In this algorithm, after the first PDA, data is clustered into c partitions using EFCM [8]. Then the data in memory is condensed into c weighted points and clustered with new points loaded in the next PDA. These points are calculated by summing the membership of examples in a cluster. In each PDA new singleton points are loaded into memory and clustered along with the past c weighted points obtained from the previous clustering. It is called Partial Data Clustering (PDC). After clustering these new singleton points along with the past c weighted points, they are condensed again into c new higher weighted points and clustered with examples loaded in the next PDA. This continues until all the data has been scanned once.

The evaluation of the clustering with WEFCM is done globally and locally. So, it uses the following measures:

Global Weighted Separation-Compactness: given a cluster scheme $C = \{C_1, C_2, \dots, C_c\}$ for a data set $X = \{x_1, x_2, \dots, x_N\}$, let $C' = \{C_{pj} | C \text{ and } C_{pj} \text{ is not singleton, } j=1,2,\dots,k\}$ where $k = |C'|$, the global weighted separation-compactness, WSC , of a cluster scheme C is given equation (1).

$$WSC = \frac{k^2}{c} \frac{\left(\sum_{j=1}^c \min_{1 \leq k \leq c} \left\{ \|v_j - v_k\|^2 \right\} / c \right)}{\sum_{j=1}^k \left(\sum_{x_i \in C_{pj}, x_i \neq v_j} w_i \mu_j(x_i)^2 \|x_i - v_j\|^2 / \sum_{x_i \in C_{pj}, x_i \neq v_j} w_i \mu_j(x_i)^2 \right)} \quad (1)$$

where $\mu_j(x_i)$ is the membership degree of data x_i to j^{th} cluster and w_i the weight of a data [19].

Then, the objective of WEFCM algorithm is to find the cluster scheme which solves:

$$\max_{2 \leq c \leq \sqrt{N}} \left\{ \max_{\Omega_c} \{WSC\} \right\} \quad (2)$$

Where Ω_c denotes all of the candidate cluster schemes for a certain number of clusters c .

We still use the measures of separation and compactness to evaluate individual clusters (except singleton).

Local Weighted Separation-Compactness: given a cluster scheme $C = \{C_1, C_2, \dots, C_c\}$ for a dataset $X = \{x_1, x_2, \dots, x_N\}$, for each $C_j \in C$, if C_j is not singleton, the local weighted separation-compactness of C_j , denoted as wsc_j , is given by equation (3) [19].

$$wsc_j = (\min_{1 \leq j \leq c, k \neq j} \|v_j - v_k\|^2) \times \left(\frac{\sum_{x_i \in C_j, x_i \neq v_j} w_i \mu_j(x_i)^2}{\sum_{x_i \in C_j, x_i \neq v_j} w_i \mu_j(x_i)^2 \|x_i - v_j\|^2} \right) \quad (3)$$

Thus, the “worst” cluster is the one with the least wsc_j value.

Procedure 1 Weighted Merge-based Process [19]

- Step 1:** Build the array $C^* = \{C_1^*, C_2^*, \dots, C_c^*, C_{c+1}^*\}$, such that each v_j^* is the center of cluster $C_j^* \in C^*$.
- Step 2:** Calculate wsc value for each C_j^* in C^* .
- Step 3:** Delete the center of cluster with the least wsc value from C^* , recalculate the cluster centers. Store the new cluster scheme $C = \{C_1, C_2, \dots, C_c\}$.

Procedure 2 Recalculate the cluster centers [19]

- Step 1:** Choose the nearest center v_j^* for each element $x_i \in X$, and group x_i into cluster C_j^* whose center is v_j^* .
- Step 2:** Calculate the data median for each C_j^* as the new center for it; denote it as v_j , output the cluster scheme $C = \{C_1, C_2, \dots, C_c\}$.
- Step 3:** If $(c^* \neq c)$ and maximum step is not reached, go to step 1.
- Step 4:** Output C .

Let p the fraction of data loaded in each PDA, d the number of PDA's required by WEFCM and $1 \leq r \leq d$.

WEFCM follows the following steps [19]:

Step 1:	The first PDA. Applying the EFCM clustering algorithm.
Step 2:	Repeat Clustering will be applied on singleton points freshly loaded in the d^{th} PDA along with c_{opt} weighted points obtained after condensation from the $(d-1)^{th}$ PDA. So, there will be $n_d + c$ points in the memory for clustering using EFCM by taking in account of weighted local and global measures. Until $r=d$.

However, data used in clustering can have hundreds attributes or variables. Clustering in such space of high dimensionality is extremely difficult. The calculation of similarity becomes very expensive.

Another problem can occur in a space of high dimension, it is the presence of non relevant attributes which affect proximity measurements. In fact, this phenomenon can appear with data of low dimension, but the possibility of presence and the number of the non relevant attributes increase with the dimensionality.

There are certain techniques for dimension reduction to solve the problems above, such as the technique of selection of attributes, the reduction of dimensionality, classification under-spaces and Co-classification [17], [15].

We present, in the next section, a new algorithm for clustering large-scale data sets while using the PCA.

III. CLUSTERING OF LARGE-SCALE DATASETS USING PCA

The clustering is often used to facilitate the training and decrease the processing time. Particularly in the case when the clustering is made on the two levels.

It should be noted that the data clustering, in order to gather them in a number of representative clusters is the most used, but we can also make a scale clustering in order to reduce their number.

A. PCA Terminology

The purpose of PCA is to create one or more variables which are a linear combination of n variables. A data described by n variables belonging to R can be represented by a point in a space R with n dimensions.

The PCA allows, starting from n variables belonging to R^n , to build p ($\leq n$) other variables, called principal components (or factorial axes), which are linear combinations of the n initial variables. Each principal component will be defined by a linear equation bringing into account the initial variables which take part in its construction.

If there are only two variables a_j and a_k , we give the graph giving a_j according to a_k and we see immediately if there is linear correlation or not. If there are more than two variables, this geometrical approach does not function any more, because we are unable to understand a space with more than 3 dimensions.

The fundamental idea of the PCA is the following one: considering the group of points in P dimensions (in this space, 1 point = 1 data), we seek to find the plan (the two-dimensional representation that we will be able to see and understand) in which the projection of the points of the cloud is least possible distorted therefore most reliable possible.

For that, we start by seeking the line Δ_1 which minimizes the squared sum of the distances between each point and the line.

This line has the property to be the direction according to which the variance of the points is biggest. Then, we look for a second line Δ_2 perpendicular to Δ_1 and have the same property; and we continue until obtaining P lines which form an orthogonal reference mark.

We note that, like any statistical method, the number of data and the number of attributes must be relatively important so that the conclusions of the PCA have a meaning;

Let take the example of data in table II which describes 10 persons by their physical characteristics and their annual average notes.

TABLE II. DATA EXAMPLE

Data	Weight	Tall	Age	Mark
1	45	1.5	13	14
2	50	1.6	13	16
3	50	1.65	13	15
4	60	1.75	15	9
5	60	1.7	14	10
6	60	1.7	14	7
7	70	1.6	14	8
8	65	1.6	13	13
9	60	1.55	15	17
10	65	1.7	14	11

B. Distances Calculation, Data Centering and Reduction

It is necessary to start by defining a distance concept between two data. So, if we take the Euclidean geometry, we could write:

$$d^2(x_i, x_j) = \sum_{k=1}^{k=P} (x_{i,k} - x_{j,k})^2 \quad (4)$$

Let us consider data x_4 , x_5 and x_6 . We have $d^2(x_4, x_5) = 2.0$, $d^2(x_4, x_6) = 5.0$ and $d^2(x_5, x_6) = 9.0$. Now if we express the tall in centimeters, the distances become: $d^2(x_4, x_5) = 27$, $d^2(x_4, x_6) = 30$ and $d^2(x_5, x_6) = 9$. It is noted that according to the unit used, data x_6 is closer to the 4th or of 5th, we use reduced centered variables to avoid this problem.

In this case, we obtain $d^2(x_4, x_5) = 2.3$, $d^2(x_4, x_6) = 2.6$ and $d^2(x_5, x_6) = 0.8$, whatever the units used.

Afterwards, we consider that the data are centered and reduced. The definition of distances given above means that all attributes have the same importance in the distance measure.

We use equation (5) to balance the attributes to have more or less importance:

$$d^2(x_1, x_2) = \sum_{k=1}^{k=P} w_k (x_{i,k} - x_{j,k}) \quad (5)$$

where a weight w_k belongs to R and balances each component of the distance for each attribute.

If all the attributes have the same importance, we take w_k all equals. We suppose here that all the attributes have the same importance and that $w_k = 1$. For this the calculation of the averages and standard deviations is essential for data centering and reduction. $\bar{x}$ is the arithmetic mean of each column, it is presented by equation (6).

$$\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i \quad (6)$$

The standard deviation is the square root of the variance; it is presented by equation (7):

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2 \quad (7)$$

Table III gives the average and the standard deviation of each variable (column).

TABLE III. CHARACTERISTICS OF THE GRAVITY CENTER

	Weight	Size	Age	Note
Average	58.5	1.64	13.8	12.0
Standard deviation	7.84	0.08	0.79	3.50

Once we calculate the average and the standard deviation of each column, we determine the centered and reduced coordinates of each data in a new reference mark whose origin is the gravity center and whose axes are parallel to the original ones (of the non reduced and non centered data).

Table IV presents the new centered and reduced data.

TABLE IV. DATA CENTRING AND REDUCTION

	Weight	Size	Age	Note
1	-1.72	-1.72	-1.01	0.57
2	-1.09	-0.45	-1.01	1.14
3	-1.09	0.19	-1.01	0.86
4	0.19	1.47	1.52	-0.86
5	0.19	0.83	0.25	-0.57
6	0.19	0.83	0.25	-1.43
7	1.47	-0.45	0.25	-1.14
8	0.83	-0.45	-1.01	0.29
9	0.19	-1.09	1.52	1.43
10	0.83	0.83	0.25	-0.29

Next, we suppose that we centered and reduced the data, $\bar{x}_j = 0$ is the average of j^{th} column and $\sigma(x_j) = 1 \forall j \in \{1, \dots, P\}$.

This assumption simplifies the calculations and the notations. The N data described by their P characters can be put in the form of a matrix having N lines and P columns: one

line describes a data and each column corresponds to an attribute. Let us note this matrix X .

C. Matrix Counting

To determine the principal components, we start from a square matrix. The choice of the matrix type is a parameter of the PCA. It is directly function of metric adapted to the problem.

- Covariance matrix

The covariance of two variables V_1 and V_2 is an indicator of the simultaneous variation. Covariance is positive when V_2 grows each time V_1 increases. It is negative when V_2 decrease each time V_1 increases. It is null if V_1 and V_2 are independent. Covariance and correlation are of the same sign. Covariance is function of the coefficient of correlation. It is given in equation (8).

$$Cov(V_1, V_2) = Stand-Dev(V_1) * Stand-Dev(V_2) * Cor(V_1, V_2) \quad (8)$$

Where Cov is the covariance, $Stand-Dev$ is the standard deviation and Cor the correlation.

If there are homogeneous data with identical magnitude orders, then there is an Euclidean metric and we work with a covariance matrix.

With each couple (V_1, V_2) , the value of each element of the matrix is that of the covariance of the couple (V_1, V_2) .

- Correlation Matrix

For each couple (V_1, V_2) , the value of each element of the correlation matrix is the correlation coefficient between V_1 and V_2 . It is given in equation (9).

$$Cor(V_1, V_2) = Cov(V_1, V_2) / (Stand-Dev(V_1) * Stand-Dev(V_2)) \quad (9)$$

If there are heterogeneous data with different magnitude orders, we work with a correlation matrix. It is given in equation (10)

$$R = \begin{pmatrix} 1 & r(a_1, a_2) & r(a_1, a_3) \dots r(a_1, a_P) \\ & 1 & r(a_2, a_3) \dots r(a_2, a_P) \\ & & \ddots & \vdots \\ & & & 1 \end{pmatrix} \quad (10)$$

R is symmetric since $r(a_j, a_k) = r(a_k, a_j)$ (the left lower triangle was not indicated). For centered and reduced variables, these two matrices are equal: $R = V$.

We have the relation: $R = \frac{1}{P} X^T X$ where X^T indicates the transposed of the matrix X . The examination of the correlation matrix is interesting: it makes it possible to immediately locate the strongly correlated variables and the variables which are not correlated.

The calculation of the eigenvalues and the eigenvectors of R provide all required information.

Let we note $\lambda_i \in \{1, \dots, P\}$ the P eigenvalues of R ordered in a decreasing way: $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_P$ and their P associated

eigenvectors $\vec{v}_i$ $i \in \{1, \dots, P\}$. R is symmetric and defined positive, its values are real and positive and its own vectors have real coordinates.

If there exists one, or many, null eigenvalues, we understand that the attributes are not linearly independent some of the others: one, or several, attributes are obtained by linear combination of the others.

Eigenvectors are unit and orthogonal two by two and form an orthonormal basis. They are the required main axes. The main axis is that associated to the biggest eigenvalue. Therefore, the q main axes are the lines whose unit vectors are the eigenvectors associated to the q biggest eigenvalues and going by G .

We obtain the principal coordinates of the data by projecting them in this space, i.e. by making the scalar product of the coordinates (centered reduced) of a data by each one of the P eigenvectors. By taking $q=2$, we can represent the N data in the principal plan.

The data distribution in the principal plan must be realized and studied. The problem is to know if the representation in the principal plan is faithful to the initial group of data. We approach this point below.

In the rest, we give an explanation on the calculation of eigenvectors and eigenvalues of the correlation matrix.

The first question is "What are we looking for"? The answer is that we seek the directions of greater lengthening of the data group, these directions having to be orthogonal to obtain an orthogonal base.

The second question is "How to determine lengthening according to a certain direction $\vec{v}$ of the data group? We determine it by the variance according to this direction which is obtained by $V^T R V$.

The third question is "How we obtain the direction of greater lengthening? We look for the vector $\vec{V}$ which maximizes this lengthening, so which cancels the derivative of this lengthening. We seek $\vec{v}$ a vector with norm 1 which cancels $\frac{d}{d\vec{v}} \frac{\vec{v}^T R \vec{v}}{\vec{v}^T \vec{v}}$. The direct calculation of this derivative don't give anything interesting, the trick consists in calculating the derivative of this same quantity divided by 1; the norm of $\vec{v}$ is 1 (by assumption); this norm is worth $\frac{\vec{v}^T R \vec{v}}{\vec{v}^T \vec{v}}$; we are interested in $\vec{v}$ which cancels $\frac{d}{d\vec{v}} \frac{(\vec{v}^T R \vec{v})/(\vec{v}^T \vec{v})}{\vec{v}^T \vec{v}}$. So that this derivative is null, the relation obtained is: $R \vec{v} = \left(\frac{\vec{v}^T R \vec{v}}{\vec{v}^T \vec{v}} \right) \vec{v} = \lambda \vec{v}$.

Therefore, $\vec{v}$ must be an eigenvector of R associated to the eigenvalue. We must be interested in the eigenvector associated

to the greatest eigenvalue. More generally, the set ordered eigenvectors according to their decreasing eigenvalues will provide the set of directions corresponding to lengthening in the decreasing order of the data group.

The matrix R is symmetric and definite positive. It has only real and positive or null eigenvalues. Moreover, eigenvectors form an orthonormal base. The rank of the eigenvalues indicates their importance. If some eigenvalues have much more important values than the others that mean that the essence of information is given by these main axes and that the others axes give little information.

On the contrary, if all the eigenvalues have approximately the same value, then no main axis can be left. The ideal situation is, when one or two eigenvalues are very important compared to the others. Notice that $\sum_{i \in \{1, \dots, P\}} \lambda_i = P$, the importance of an eigenvalue compared to the others is well measured by its inertia. It is given in equation (11).

$$\lambda_i = \frac{\lambda_i}{P} \quad (11)$$

λ_i is the variance along the main axis carried by the eigenvector; therefore $\vec{V}_i$ is the variance of the principal component z_i .

Also we note that the covariance matrix in principal space is diagonal: by a PCA, we obtain a representation in a data space in which the components are not linearly correlated. It is a very interesting property, knowing that projection in the principal plan deforms reality by shortening the distances (the distribution of the data in the first space).

It is necessary to be able to estimate this deformation. For that, we calculate the cosine of the angle formed by the half-line going of the origin to the initial point and the half-line going of the origin to the projection of the point in principal space, either an axis, or a plan.

We note $\theta_{i \in \{1, \dots, N\}, j \in \{1, \dots, p\}}$ the angle between the half-line passing by the origin and the initial point x_i and the main axis j . $\theta_{i \in \{1, \dots, N\}, j \in \{1, \dots, p\} \times k \in \{1, \dots, p\}}$ is the angle between the half-line passing by the origin and the plan formed by the main axes j and k . The square cosine of these angles is given in equation (12) and (13).

$$\cos^2 \theta_{i \in \{1, \dots, N\}, j \in \{1, \dots, p\}} = \frac{z_{ij}^2}{\sum_{j \in \{1, \dots, p\}} x_{ij}^2} \quad (12)$$

$$\cos^2 \theta_{i \in \{1, \dots, N\}, j \in \{1, \dots, p\} \times k \in \{1, \dots, p\}} = \frac{z_{ij}^2 + z_{ik}^2}{\sum_{j \in \{1, \dots, p\}} x_{ij}^2} \quad (13)$$

Where z_{ij} and z_{ik} are the coordinates on the main axes j and k of the data x_i . A square cosine lower than 0.3 indicates an important deformation ($>57^\circ$).

The linear coefficient of correlation between the j^{th} initial variable x_j and the k^{th} principal component z_k is given in equation (14).

$$r_{j,k} = v_{k[j]} \times \sqrt{\lambda_k} \quad (14)$$

All the axes can have their utility. To judge some, it is necessary to compare the various eigenvalues (seen above). For each one, we can measure their inertia I_j and their cumulated inertia. It is given in equation (15).

$$C_i = \sum_{j \leq i} I_j \quad (15)$$

D. Dimension Reduction

The dimension reduction procedure makes it possible to reduce the number of variables of the data set while maximizing the information (variance) contained in the simplified model.

Procedure 3 Dimension Reduction by PCA

- Step 1:** Data centering and reduction by using equations 6 and 7.
- Step 2:** Calculation of the correlation or covariance matrix using the equations 8 and 9.
- Step 3:** Calculation of eigenvalues and eigenvectors of the matrix.
- Step 4:** Sort out the eigenvalues thanks to their inertia given by the equation 11.
- Step 5:** Select the $K \leq P$ eigenvectors and generate the whole data in a new presentations structure using the equation 15.
- Step 6:** Look into the correlation between the initial and the principal data using the equation 14.

This procedure will be applied to the data set in order to ensure a dimension reduction. Once reduced, WEFCM can be applied to classify the data. The new algorithm was named SR-WEFCM (Scale Reduction-Weighted Enhance Fuzzy C-Means).

E. Illustrative Example

For a better assimilation of PCA, we take again the example given above with persons and their characteristics. It is given by equation (16).

$$R = \begin{pmatrix} & \text{Weight} & \text{Tall} & \text{Age} & \text{Mark} \\ \text{Weight} & 1 & 0.37 & 0.48 & -0.56 \\ \text{Tall} & 0.36 & 1 & 0.39 & -0.32 \\ \text{Age} & 0.48 & 0.39 & 1 & -0.32 \\ \text{Mark} & -0.56 & -0.62 & -0.32 & 1 \end{pmatrix} \quad (16)$$

The eigenvalues and eigenvectors are given in equation (17).

$$\lambda_1 = 2.391, \quad \vec{v}_1 = (0.5080, 0.5038, 0.4455, -0.5383) \quad (17)$$

$$\lambda_2 = 0.750, \quad \vec{v}_2 = (0.3065, 0.4641, 0.7085, -0.4381)$$

$$\lambda_3 = 0.584, \quad \vec{v}_3 = (-0.6593, 0.5253, 0.4712, -0.2593)$$

$$\lambda_4 = 0.274, \quad \vec{v}_4 = (-0.4619, 0.5042, 0.2855, -0.6715)$$

We notice that no eigenvalue is null and we can verify that the 4 eigenvectors are unit and orthogonal two to two. We can calculate the inertia and the cumulated inertia of each eigenvalue in table V.

TABLE V. INERTIA AND CUMULATED INERTIA FOR EACH EIGENVALUE

λ	I	C
2.391	0.59775	0.59775
0.750	0.1875	0.78525
0.584	0.146	0.93125
0.274	0.0685	1.0

The data coordinates in the new reference mark are calculated as given in equation (18).

$$\begin{pmatrix} (x_{1,1} & x_{1,2} & x_{1,3} & x_{1,4}) \vec{v}_1 \\ (x_{1,1} & x_{1,2} & x_{1,3} & x_{1,4}) \vec{v}_2 \\ (x_{1,1} & x_{1,2} & x_{1,3} & x_{1,4}) \vec{v}_3 \\ (x_{1,1} & x_{1,2} & x_{1,3} & x_{1,4}) \vec{v}_4 \end{pmatrix} \quad (18)$$

For example, the coordinates of first data are: (-2,64;-0,20;-0,10;1,04).

The data coordinates on the principal's axis are given in table VI.

TABLE VI. DATA COORDINATES IN THE NEW PRINCIPAL PLAN

Data	Coordinates
1	(-2.64, -0.20, -0.10, 1.04)
2	(-1.94, -0.36, 0.32, -0.35)
3	(-1.44, -0.80, 0.59, -0.35)
4	(2.08, -0.08, 1.20, 0.19)
5	(0.99, -0.42, 0.30, -0.05)
6	(1.47, -0.82, 0.06, 0.56)
7	(1.32, -0.35, -1.45, 0.41)
8	(-0.43, -0.14, -1.25, -0.67)
9	(-0.57, 2.39, 0.41, -0.07)
10	(1.17, -0.08, -0.07, -0.57)

In table VII, we calculate the correlation between initial variables and principal components.

TABLE VII. STUDY OF THE CORRELATION BETWEEN INITIAL VARIABLES AND PRINCIPAL COMPONENTS

	z_1	z_2	z_3	z_4
Weight	0.785	0.266	-0.504	-0.242
Tall	0.779	-0.403	0.402	-0.264
Age	0.689	0.611	0.360	0.150
Mark	-0.832	0.380	0.198	-0.352

We notice that: the 4 variables are strongly correlated (in absolute value) with the first main axis. The variable Mark is correlated negatively with the three other variables. And finally to estimate the deformation due to projection in the principal plan, we calculate $\cos^2$. It is given in table VIII.

TABLE VIII. ILLUSTRATION OF THE DEFORMATION DUE TO THE PROJECTION

Data	$\cos^2 \theta_{1,1}$	$\cos^2 \theta_{1,2}$	$\cos^2 \theta_{1,3}$	$\cos^2 \theta_{1,4}$
1	0.859	0.005	0.82	0.82
2	0.916	0.031	0.94	0.94
3	0.628	0.195	0.83	0.74
4	0.745	0.001	0.75	0.99
5	0.785	0.142	0.88	0.79
6	0.690	0.212	0.87	0.67
7	0.419	0.030	0.45	0.93
8	0.084	0.008	0.09	0.79
9	0.053	0.919	0.97	0.08
10	0.804	0.004	0.80	0.80

IV. EXPERIMENTATION AND EVALUATION

We evaluated in experiments the algorithm SR-WEFCM on large-scale data such as isolet5. We then carried out a comparison of the results given when we applied only WEFCM to the data and in the case when we applied SR-WEFCM.

ISOLET5 is a subset of the ISOLET (Isolated Letter Speech Recognition) set, the letter of the experiment of voice recognition. This subset was prepared in the same way that in [18]. Ten classes out of 26 (represents the letters of the alphabet) were selected by chance (this set is composed of 1559 data and 617 attributes) [18]. The optimal number of clusters for this whole data is 10.

The comparative study will be made according to tow axes: the first consists of the dimension reduction, the second will be the clustering quality.

A. Dimension Reduction

Our dimension reduction procedure using PCA produces two files: the first contains a general data analysis like columns average, their standard deviation, the values, the vectors and some other information about correlation between the initial variable and principles components. The second file contains new coordinates of each data in the new plan.

The performance and importance of the dimension reduction approach using PCA becomes clear when we examine the isolet5 data set. Thanks to this method we could reduce the variables number. After this analysis, we obtain 3 principles components instead of 617 attributes. Each principle component presents and synthesizes a set of initial variables.

During this first step, we could reduce the attributes number to a restricted number of principles components. In other word, we were aware of information quantity observed in the initial data set while tacking only few reduced principles components. We obtained the new coordinates in the new mark. The next step will be the data reduction, which means

that we will apply the WEFCM clustering algorithm on the new data set.

B. Clustering Quality

We applied WEFCM and SR-WEFCM to ISOLET5, 10 is the optimal the number of clusters. The matrix of the output centers was the following one:

TABLE IX. MATRIX OF CENTERS BY WEFCM

	Variable 1	...	Variable 100	...
Cluster 1	-0.032180	...	0.103958	...
...	...	...	...	...
...	...	...	...	...
Cluster 10	0.074184	...	0.055395	...

We point out that each element contains 617 Attributes, and therefore we chose to represent only some attributes, and only for some clusters.

TABLE X. MATRIX OF GENERATED C ENTERS BY SR-WEFCM

	Variable 1	Variable 2	Variable 3
Cluster 1	0.216584	-0.036261	-0.000480
...	...	...	...
...	...	...	...
Cluster 10	-0.002041	0.000196	0.152780

In the same way, we drew the tables of the associated indices relating to ISOLET5.

Table VII presents the validity indices obtained when we use WEFCM clustering algorithm.

TABLE XI. VALIDITY INDICES WITH WEFCM

Number of clusters	Validity indices
11	0.026344
10	0.050870
9	0.046293
8	0.028907
7	0.029964
6	0.020478
5	0.017594
4	0.013673
3	0.006801
2	0.001851

As we can see, the optimal number of clusters is obtained with the biggest value of validity index.

Table VIII presents the validity indices obtained when we use SR-WEFCM clustering algorithm.

TABLE XII. VALIDITY INDICES WITH SR-WEFCM

Number of clusters	Validity indices
11	0.049331
10	0.283013
9	0.030206
8	0.044234
7	0.036985
6	0.245792
5	0.120833
4	0.022860
3	0.046677
2	0.001674

According to table VIII, the biggest value of the index corresponded to the optimal number of clusters is equal to 10.

We note that in spite of the dimension reduction, the suggested system proved its performances in the determination and the evaluation of the number of clustering.

Table IX presents the run time of both WEFCM and SR-WEFCM clustering algorithms.

TABLE XIII. RUN TIME

WEFCM	71.076442
SR-WEFCM	0.288956

The importance of our system was quite remarkable also when we examine the run time of the clustering algorithm for large-scale data set, (the isolet5 in our case). It is noticed that the difference in run time is very important, so, we can say that the reduction of large-scale data set by dimension reduction procedure involves necessarily a run time reduction.

V. CONCLUSION

Recent progress in data filing field offers new challenges to the theorists of Data Mining. One of these challenges is to manage large data sets, in which traditional algorithms and researcher's intuition have multiple difficulties.

In this direction, clustering proposes important and effective techniques of data sampling and compression in order to reduce the memory size dedicated to the data which will be classified.

So, it is possible to know, that only a small number of variables have interest. In this case, these variables can be selected and the other thrown, in order to reduce the data set dimensionality. More generally, dimension and data reduction prove to be necessary for the management of large-scale data sets.

The scale-reduction is a complex problem which has been largely studied. In this paper, we have proposed a new clustering algorithm, called SR-WEFCM, used in the case of large-scale data sets and based on, both a compression technique to reduce the data, and PCA to reduce the variables (or attributes).

With SR-WEFCM, our objective is reached. We work out an automatic clustering algorithm for large-scale data set. This is ensured through both an enhanced compression technique and validity measures which allow the evaluation of obtained clustering while giving the optimal number of clusters. It ensures a double reduction, the dimension and the data.

As future work, we propose to i) test our approaches with other large-scale data sets and ii) take into account uncertain data and keep only the data representative by using a selection variables technique.

REFERENCES

- [1] J. Han and M. Kamber, Data Mining: Concepts and Techniques, San Francisco:Morgan Kaufmann Publishers, pp.105-130, 2001.
- [2] T. Zhang, R. Ramakrishnan, and, M. Livny, "BIRCH: An efficient data clustering algorithm for large databases," ACM SIGMOD International Conference on Management of Data, pp. 73-84, 98, Montreal, Canada, 1996.
- [3] J. McMueen, "Some methods for classification and analysis of multivariate observations," The Fifth Berkeley Symposium on Mathematical Statistics and Probability, pp. 81-297, 1967.
- [4] J. Bezdek, Fuzzy mathematics in pattern classification, Ph.D. Dissertation, Cor-nell University, 1973.
- [5] W.P.Vogt, Dictionary of statistics and methodology, A nontechnical guide for the social sciences, CA: Sage, pp. 177,1993.
- [6] J. Bezdek, Chapter F6: Pattern Recognition in Handbook of Fuzzy Computation, IOP Publishing Ltd, 1998.
- [7] H.Sun, "A theory on distinguishing overlapping components in mixture models," Research Report, DMI, University of Sherbrooke, n° 345, November 2003.
- [8] M. Sassi, G.Touzi, and H. Ounelli, "Using Gaussians Functions to Determine Representative Clustering Prototypes," 17th IEEE International Conference on Database and Expert Systems Applications, Poland, pp. 435-439, 2006.
- [9] P. Hore, L. Hall, and D. Goldgof, Single Pass Fuzzy C Means, IEEE International conference on Fuzzy Systems, pp. 847-852, 2007.
- [10] P. Hore, L.Hall, and D.Goldgof, "A Cluster Ensemble Framework for Large Datasets," IEEE International Conference Systems, Manand Cybernetics, 2006.
- [11] P.S Bradley and U.M. Fayyad, "Refining initial points for k-means clustering," 15th International Conference on Machine Learning, Morgan kaufmann, San Francisco, pp. 91-99, 1998.
- [12] J. Vaillant, Notion de base en échantillonnage, Université des Antilles-guyanne, 1996.
- [13] M. Aounallah, S. Quirion and G. Mineau, Forage distribué des données: une comparaison entre l'agrégation d'échantillons et l'agrégation de règles, 2004.
- [14] P.S. Bradley, U.M. Fayyad and C.A. Reina, Scaling EM (Expectation-Maximization) Clustering to Large DataBases, October 1998.
- [15] D. P. Mercer, Clustering large datasets. Technical report, Linacre College, Octobre 2003.
- [16] C.Candillier,thesis extract: " Méthodes d'Extraction de Connaissances à partir de Données (ECD) appliquées aux Systèmes d'Information Géographiques (SIG)," UFR Sciences et techniques, Université de Nante, September 2006.
- [17] T.Ng. Raymond and J.Han, "CLARANS: A Method for Clustering Objects for Spatial Data Mining," IEEE Transactions on knowledge and data engineering, vol. 14, no. 5, September/October 2002.
- [18] Machine Learning Repository . Available at: <http://archive.ics.uci.edu/ml/>.
- [19] M. Sassi, A. Grissa, "Clustering Large Data sets based on Data Compression Technique," IEEE conference on Fuzzy Systems (FUZZ-IEEE), Korea, August 20-24, pp. 396-402, 2009.