International Journal of Electrical and Computer Engineering (IJECE)
Vol. 11, No. 3, June 2021, pp. 2109~2119
ISSN: 2088-8708, DOI: 10.11591/ijece.v11i3.pp2109-2119

Clustering using kernel entropy principal component analysis and variable kernel estimator

Loubna El Fattahi[1], El Hassan Sbai[2]
[1]Department of Physics, Moulay Ismail University of Meknes, Morocco
[2]High School of Technology, Moulay Ismail University of Meknes, Morocco

Article Info

Article history:

Received Dec 9, 2019
Revised Sep 26, 2020
Accepted Oct 6, 2020

Keywords:

Clustering
Kernel entropy principal component analysis
Maximum entropy principle
density peak
Variable kernel estimator

ABSTRACT

Clustering as unsupervised learning method is the mission of dividing data objects into clusters with common characteristics. In the present paper, we introduce an enhanced technique of the existing EPCA data transformation method. Incorporating the kernel function into the EPCA, the input space can be mapped implicitly into a high-dimensional of feature space. Then, the Shannon's entropy estimated via the inertia provided by the contribution of every mapped object in data is the key measure to determine the optimal extracted features space. Our proposed method performs very well the clustering algorithm of the fast search of clusters' centers based on the local densities' computing. Experimental results disclose that the approach is feasible and efficient on the performance query.

Corresponding Author:

Loubna El Fattahi
Department of Physics
Moulay Ismail University of Meknes
Km 5, Rue d'Agouray ‹N6, Meknes 50040, Morocco
Email: estm@est-umi.ac.ma

1. INTRODUCTION

Clustering the data is the task of discovering natural groupings in data, named clusters according to their similarity. Objects are similar inside the same cluster whereas dissimilar compared to objects descending from other clusters. Clustering, as a class of unsupervised classification method, has been widely applied in different domains, machine learning, image segmentation, pattern recognition, text mining and many other domains [1-3]. Great number of clustering algorithms lie in literature, the famous K-mean clustering [4], hierarchical clustering [5], k-medoids [6], and mean shift [7] have been considered in various problems.

Despite of the extensive studies in the past on clustering, a critical issue remains largely unsolved: how to automatically determine the number of clusters. Most of them assumed that the number of clusters either has been manually set or is known in advance [4, 8]. Recently, great attention has been accorded to tackle with this issue. Clustering by density peaks selection criterion [9-11] is a technique that tends to find intuitively the number of clusters independently of their shape and the dimension of the space containing them, so that to avoid the inherent shortcomings of providing the number of clusters as an input parameter like in the K-means. The approach proposed by Rodriguez and Laio [10] relies as a first step on the fast search of the density peaks referring the cluster centers characterized by having a higher local density compared to their neighborhood and relatively large distance regard to points having higher densities. As a result, the cluster centers are located in general at the upper right corner of the decision graph which is the plot of the density as a function of the distance of each point. Therefore, once centers are determined all

remaining data points are assigned to the same cluster as its nearest neighbors of higher density so as to form final clusters.

As a further work, the Rodriguez and Laio's method was improved by Wang and Xu [11] through substituting the step function by the Parzen estimator so that the truncated density becomes smoother. Here, we develop a method for data transformation and reduction based on the variable kernel estimator [12]. Aimed at the problems of clustering, many researchers are convinced that the dimensionality reduction is an important stage that must be adopted in data analysis before considering any classification step. In fact, many algorithms of clustering often do not work well in high dimension, so, to improve the efficiency, a data reduction is needed [1]. In this sense, we propose a hybrid method, which combines the entropy principal component analysis (EPCA) [13, 14] and the data mapping. The core element is to perform a data mapping using the kernel function before implementing the EPCA. Data mapping consists of transforming the data into a high-dimensional feature space, where patterns become linear and the nonlinearity disappears [15]. Then, using EPCA, we restrict the high-dimensional space to a subspace of the extracted features.

In many cases, the quality of clustering is approved by a low similarity of the inter-cluster and a high similarity of the intra-cluster. To this end, we integrate an automatic method for cluster centroid selection based on the validity index algorithm using the concept of entropy introduced by Jayens Edwin in [16] and computing the inter-cluster and intra-cluster similarities [17].

The present paper is organized as follows. Section 2 presents a brief review of PCA as a linear data transformation. We move to consider the kernel entropy principal component analysis (KEPCA) for dimensionality reduction. In section 3, we present the clustering by the fast search of centers relying on the estimation of the probability density function (PDF) as well as the automatic criterion for the selection of cluster centers using validity index algorithm. Section 4, is dedicated to the validation of the proposed method on both synthetic and real datasets including a comparison to other clustering algorithms.

2. RESEARCH METHOD

2.1. Dimensionality reduction

Clustering algorithms are facing problems of dimensionality especially when the dimension increases importantly. Therefore, in some cases, they lose their efficiency, likewise, their productivity especially when data present sparseness. As a solution, we consider the reduction of the dimensionality as an efficient preprocessing. Due to this effect, many researches have been done to get rid of this complication [18-21]. Thus, we establish a nonlinear method named KEPCA, which improves the existed linear EPCA method [13]. The purpose is to discard the redundant and irrelevant information and get only the valuable one. Using the maximum entropy principle [16], we can easily determine the reduced dimension of data in kernel space.

2.1.1. Principal component analysis

Principal component analysis (PCA) is very famous as a technique of multivariate statistics. Due to the fact of analyzing data in term of feature extraction and dimensionality reduction, PCA is adopted by almost all disciplines. The ultimate objective of PCA is to select variables from input data table which have higher statistical information then squeeze out this information as a set of new orthogonal variables called principal component based on mathematic notions: eigenvalues, eigenvectors, mean and standard deviation [14, 20].

2.1.2. Shannon entropy

Claude Shannon established the entropy concept in information theory. He introduced the term $-log\ ((X_i))$ as a measurement of the information carried by the realization X_i knowing the probability of distribution p of the discrete variable X [22]. Relating to a discrete variable, the entropy measures the uncertainty. The Shannon entropy related to X is obtained calculating the following formula (1):

$$S(p) = -\sum_{i=1}^{n} p(X_i)\, log\, p(X_i) \quad (1)$$

where $p = \{p_1, p_2, \dots, p_n\}$. Thus, by combining the Shannon entropy and the PCA it gives the EPCA, where the core element is maximum entropy principle (MEP) [17] so as to determine the optimal dimension of the principal subspace keeping the maximum information.

2.1.3. Kernel entropy principal component analysis

Our proposed approach, KEPCA, is a nonlinear version of EPCA [23]. The basic aspect of kernel EPCA method is to map input data $x_t = x_1, \dots, x_N$ such that $t = 1, .., N$ into kernel space through the kernel function. Then, kernel matrix is given by $\Phi: \Re^d \to \mathcal{F}$ where $x_t = \Phi(x_t)$, is $\Phi = [\phi(x_1), \dots, \phi(x_N)]$. As soon as data mapping is done, EPCA is implemented in F. The positive semi-definite kernel function provides data mapping, $k_\sigma = R^d \times R^d \to R$ which produces an inner product in the Hilbert space F:

$$k_\sigma(x_t, x_{t'}) = \langle \phi(x_t), \phi(x_{t'}), \rangle \quad (2)$$

where every single element (t, t'), of the (N, N) kernel matrix K, is equivalent to $k_\sigma(x_i, x_{t'})$. Thus, the inner product is $K = \Phi^T \times \Phi$. To elucidate more explicitly how we proceed in implementing the EPCA, let $X_1, \dots, X_n$ be the mapped data contained in K defined by n features $f_1, \dots f_n$, and E_q is the subspace with $q = 1, \dots, n$.

The average information supplied by the contribution of each mapped element to the construction of the subspace of projection is the explained inertia, whereas the average information given by the contribution of every mapped individual to the loss of inertia is the residual inertia. Both contributions to the explained inertia (EIC) and to the residual inertia (RIC) of each single individual X_i all over the subspace of features E_q are successively given as a probability distribution in (3) and (4) [23].

$$p_q^1(X_i) = EIC(X_i, E_q) \quad (3)$$

$$p_q^2(X_i) = RIC(X_i, E_q) \quad (4)$$

with $\sum_{i=1}^n EIC(X_i, E_q) = 1$, and $\sum_{i=1}^n RIC(X_i, E_q) = 1$. Thus, the Shannon entropy provided by these distributions respectively is given asin (5) and (6):

$$S_1(p_q^1) = -\sum_{i=1}^{n} p_q^1(X_i)\ log\ p_q^1(X_i) \quad (5)$$

$$S_2(p_q^2) = -\sum_{i=1}^{n} p_q^2(X_i)\ log\ p_q^2(X_i) \quad (6)$$

The variation of the quantities in (5) and (6) is antagonist. According to the maximum entropy principle, the maximized sum of the both entropies of the probability distributions corresponds to the minimum dimension of the subspace of features.

$$S(q^*) = -max(S_1(p_q^1) + S_2(p_q^2) \quad (7)$$

q^* is the optimal dimension of the feature subspace.

2.2. Clustering by density peak selection

Recent researches give big interest to the clustering by the fast search of cluster centroids based on the nonparametric estimation of the PDF. The extended version of Rodriguez and Laio's method [10] given by Wang and Xu [11] is summarized here through the Parzen estimator rather than the step function. The main aspect of the method is the measurement of the couple (density, distance) values characterizing each data point.

2.2.1. Density estimation and distance

Relying on Rodriguez and Laio's clustering method [10], we resume here the computation of the density and distance values. The main role is to identify the cluster centroids based on the two assumptions: the first one establishes the fact that each cluster center is enclosed by elements, which have local densities lower than the center density. The Second one esteems that each cluster center is far enough from all other points with higher density. For a given data point X_i, the local density and the distance are defined respectively as (8) and (9):

$$\rho_i = \sum_{j=1}^{n} I\big(d(X_i, X_j) < d_c\big), \tag{8}$$

$$\hat{\delta}(x_i) = \min_{j:\hat{f}(x_i)<\hat{f}(x_j)} d(x_i, x_j). \tag{9}$$

where $I(A)$ is the step function of the set A; $d(X_i, X_j)$ is the Euclidian distance between two different data points; and d_c is the cut-off distance defined in advance. Analyzing definition (8), we could understand that ρ_i is simply the count of points that are closer than d_c to the ith data point, whereas in (9), the measurement δ is determined as the minimum distance among distances computed between the ith data point and all other points having higher density. Finally, the point, which has the highest density, δ_i is defined to be $max_j\ d(x_i, x_j)$.

Nonetheless, the choice of d_c is not always useful because the result of the algorithm fundamentally depends on it. The cause is that d_c describes the average number of neighbors, close to 1% and 2% of the whole number of data points. Consequently, the choice of d_c is unsystematic and unstable when the size of the sample is changing. To get rid of the bad effect of d_c, we opt using the Parzen estimator and the variable kernel estimator in lieu of the step function that has good effect on the query of performance [10]. Concerning the bandwidth parameter, it is efficiently computed using the rule of Silverman [24].

2.2.2. Variable kernel estimator

The variable kernel estimator (VKE) is a combination of Parzen estimator where the scale of the bumps placed on the data points are allowed to vary from data point to another [25, 26] and the kNN estimator.

The estimator of Parzen-Rosenblatt is defined as in (10:

$$\hat{f}(x) = \frac{1}{nh^d} \sum_{i=1}^{n} k_d \left(\frac{D(x, X_i)}{h}\right) \tag{10}$$

where K_d is the kernel, h is the smoothing parameter and $D(x, X_i)$ is the Euclidean distance between x and X_i. The k-nearest neighbors (kNN) estimator is defined as in (11) [24]:

$$\hat{f}_{knn}(x) = \frac{(k\ /\ n)}{(V_k(x))} = \frac{k/n}{c_d r_k(x)} \tag{11}$$

where k is a positive integer, $r_k(x)$ is the distance from x to the kth nearest point and $V_k(x)$ is the volume of a sphere of radius $r_k(x)$ and $c_d$is the volume of the unit sphere in d dimensions. The smoothness degree of this estimator is affected by the parameter k, taken to be very smaller than the sample size. The VKE is constructed similarly to the classical kernel estimator. It is defined by (12) [24]:

$$\hat{f}(x) = \frac{1}{nh^d} \sum_{i=1}^{n} \frac{1}{(r_{i,k})^d} k_d \left(\frac{D(x, X_i)}{h r_{i,k}}\right) \tag{12}$$

with $r_{i,k}$ is a Euclidean distance between a data point X_i and the kth nearest point of the other $n-1$ data points.

2.2.3. Cluster validity index V_{MEP}

For every clustering process, a group of clusters $c_1, \dots, c_j, \dots, c_k$ is obtained from a given dataset. The measurement P_{ij} is the relation between each point i and the cluster c_j, for $j = 1, \dots, k$. For all the pre-defined clusters c_j, we set $P_{ij} = 0$ in case $i \notin c_j$ and, when $i \in c_j, P_{ij} > 0$, we have [12, 17, 23]:

$$\sum_{i \in C_j} P_{ij} = 1, \quad for\ j = 1, \dots, \mathrm{k} \tag{13}$$

Each class provides information which is measured using the entropy formula (14):

$$S_j = -\sum_{j=1}^{k} P_{ij}\, log(P_{ij}) \quad (14)$$

Finally, the V_{MEP} index is recognized as an entropy by (15):

$$V_{MEP} = S = \frac{1}{k}\sum_{j=1}^{k} S_j + log(k^*) \quad (15)$$

where S_j is an entropy and k^* is the optimal number of classes for which the entropy S is maximal.

2.2.4. Algorithm of the Kernel entropy principal component analysis

According to the kernel method, the input space can be indirectly mapped into a high-dimensional feature space through which the nonlinearity could be removed or reduced. The GRB Function is the principal element of the kernel function because it is typically used in Reproducing Kernel Hilbert Space (RKHS) with the objective of maximizing the feature space variance of the output variables. Subsequently, the mapped data were reduced using the EPCA as a linear method for data reduction with the ultimate objective to maintain features expected to preserve as possible the valuable information. Eventually, a simple algorithm of clustering would be able to achieve significant results based on both Parzen estimator and variable kernel estimator.

Our proposed algorithm of clustering considering our proposed KEPCA method, is resumed in the next steps.

- Normalize the input data ;
- Map the input data;
- Perform the EPCA;
- Reduce the transformed data;
- Compute the bandwidth ;
- Compute the pair density-distance of reduced data;
- For $C = 1, \dots, C_{max}$
  - Select the cluster peaks and assign each element into its cluster;
  - Compute the cluster validity index $V_{MEP;}$
- The correct number of clusters and the best grouping for the input data corresponds to the one that have maximum value of V_{MEP}.

3. RESULTS AND DISCUSSION

Our present section has as concern, demonstrating the performance query of our approach by reducing data and extracting only the valuable information. Inspired by computing the couple density-distance values through the fast search of cluster centers. The use of either the Parzen estimator [27] or variable kernel estimator integrating the rule of Silverman has good outcomes on our clustering results. A comparison between the results using our clustering algorithm based on KEPCA data transformation and other algorithms of clustering is given. The artificial, the real well-known (Iris, Seeds, Flame and Heart) datasets were used [28] as well as a vehicle trajectory dataset. Our analysis study is demonstrated on MATLAB environment.

3.1. Simulated study

We consider as a first application, a simulated data that consists of three random clusters laid in 600 by 2 matrix. All clusters were generated by a normal distribution but with different random covariate matrices and centers. Cluster contains 200 data points for each. Figure 1, reveals the plot of the initial data of the two variables of the input space before considering our clustering algorithm.

As it is shown in Figure 2(a), we can observe that after executing the KEPCA data transformation result has given three dimensions as the reduced dimension after data mapping. Thus, for synthetic data, three-dimensional array are enough to represent input data in feature space. The Figure 2(b) discloses that the centers are identified as three. They are located at the upper-right corner of the density-distance plot, discriminated in red color, and circled shape. Besides, on the Figure 2(c) it can be easily understood that the sorted quantity (product of density and distance) has high value for the first three data points. This magnitude is by its own definition large, starts increasing abnormally between cluster centers and getting very narrow between other samples. Consequently, both techniques show the same number of centers, which eventually confirms the existence of three clusters. The obtained results can be explained thanks to the fact that each

center is recognized by higher local density and relatively large distance away from the other data points with higher density. Hence, our algorithm is capable of differentiating amongst centers and other data points. Next, to examine the validity of our clustering result, we investigate both V_{MEP} criterion and the Elbow method [29]. The given results are displayed on the same Figure 2. On the Figure 2(d), we can observe the progress of the V_{MEP} index through which the optimal number of clusters is identified to be three clusters applying the maximum entropy principle. Likewise, same result was given on Figure 2(e) by investigating the Elbow method, which relies on the minimization of the sum of the squared errors within each cluster. Three clusters were picked. On the last Figure 2(f) we can see the samples assignation to their convenient clusters considering the center selection outcome. Every single point is assigned to the nearest center based on the Euclidean distance calculation. Eventually, the result of the clustering of the proposed algorithm is given on three-dimensional feature space plot for the synthetic dataset since the dimension was reduced into three features after performing KEPCA Figure 2(a). The three clusters were properly distinguished one from another by different shapes and colors as the best grouping for samples, which demonstrates our clustering algorithm assumptions.

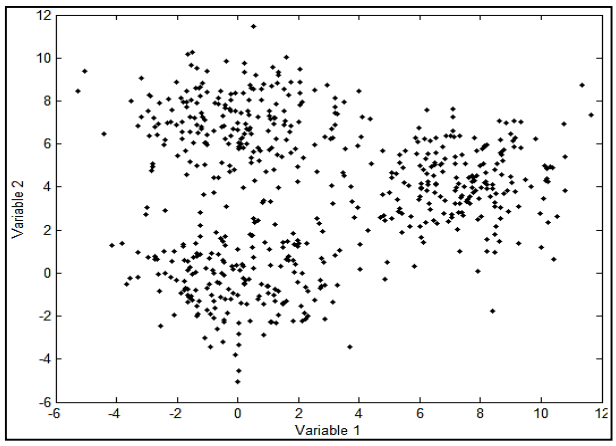

Figure 1. Two-dimensional plot of the initial artificial dataset

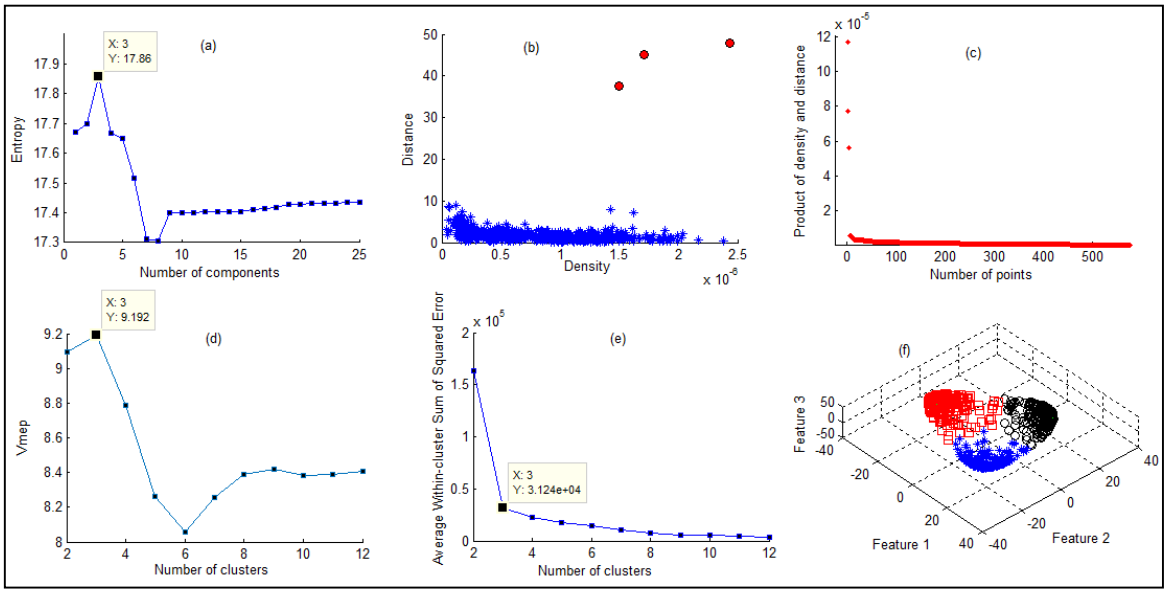

Figure 1. Results obtained on artificial dataset (a) Result of the optimal components number using the KEPCA, (b) decision graph for centers selection of the density distance plot, (c) the product of the density and distance plotted in decreasing order, (d) number of clusters validation given by V_{MEP} index, (e) the Elbow method for artificial dataset, (f) the assignment of samples to clusters

The classification rate of our proposed algorithm with KEPCA data transformation and the VKE is about 97.17% for the simulated dataset, whereas the K-means algorithm achieved 96.83%. Then, using our algorithm of clustering with the VKE and EPCA has given 82.33%. Therefore, the present algorithm with the kernel data transformation has given relatively higher classification rate than the other clustering algorithms.

3.2. The Iris dataset

The Iris benchmark is one of the machine-learning datasets; Fisher first used it in [30]. It consists of 150 measurements of three distinct types of the Iris plant (Iris setosa, Iris virginica and Iris versicolor) of the four variables: width and length for sepal and petal. It is worth mentioning that, one of the classes is linearly separable, whereas the two others are not [30]. Considering the combination of kernel data transformation (mapping) and the EPCA as a preprocessing for input data, the figure presents results obtained on Iris dataset, Figure 3(a) illustrates the reduced features in the high-dimensional space (150 features). We restrict our plot to only 25 features on the Figure 3 for the sake of clearness as the entropy evolves in decreasing order. Therefore, the seventeen maintained nonlinear features interpret the more relevant ones for our clustering algorithm in term of information content. On the Figures 3(b) and 3(c) both graphs are given relying on the measurement of the pair density and distance. Considering the first plot on the Figure 3(b), it presents a density-distance plot, whereas the Figure 3(c) it presents the sorted quantity of the density and distance product. Thus, identical result was given by both techniques, three centroids have been determined from data. The obtained results are interpreted thanks to the coming assumption: Every center is distinguished from the other data points by its higher local density and relatively large distance. Hence, our algorithm is capable of differentiating among centers and the other data points. To demonstrate our clustering algorithm performance, we incorporate the cluster validity index founded the maximum entropy principle in Figure 3(d) and the Elbow method in Figure 3(e). Both criteria perform same result, which consequently confirms the existence of three clusters.

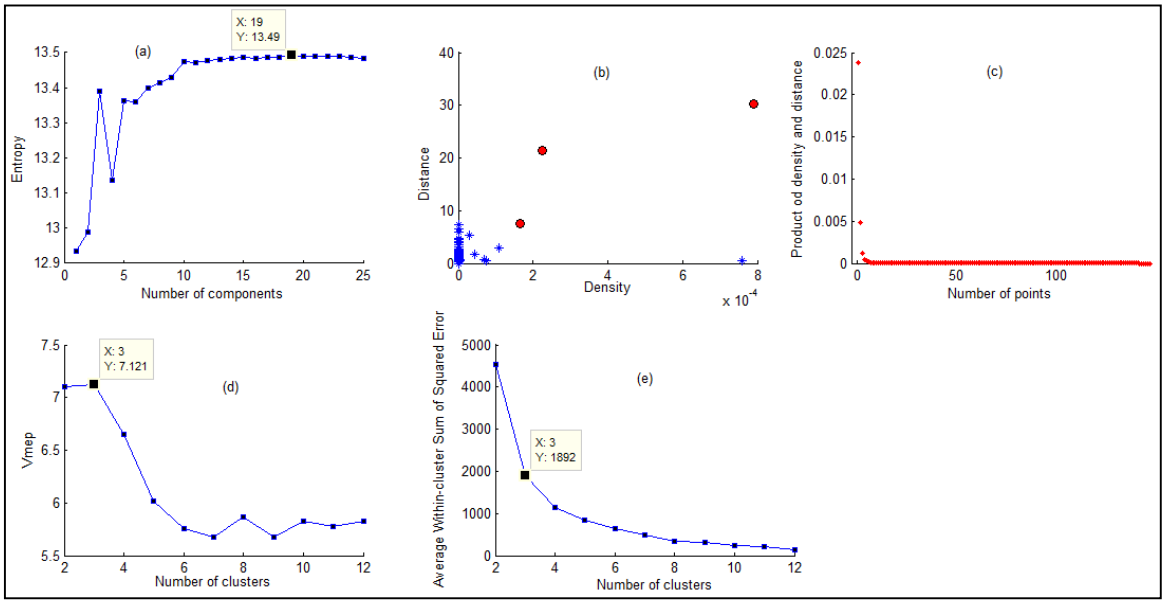

Figure 2. Results obtained on Iris dataset (a) result of the optimal components number using the KEPCA, (b) decision graph for centers selection of the density distance plot, (c) the product of the density and distance plotted in decreasing order, (d) number of clusters validation given by V_{MEP} index and (e) the Elbow method for Iris dataset

Therefore, our clustering algorithm is able to identify automatically the center of each cluster. The classification rate for Iris data using our algorithm of clustering considering the KEPCA data transformation as a preprocessing step is 89.33% incorporating variable kernel estimator while it is equals to 88% incorporating Parzen estimator, for the EPCA data transformation integrating VKE, it is unable to detect all three clusters and considers only 2 clusters, likewise for K-medoids. Then, the K-means has reached only 82%. Therefore, our automatic algorithm integrating KEPCA data transformation and the variable kernel estimator is capable of classifying patterns with a higher classification rate compared to the other algorithms.

3.3. Seeds database

The Seeds dataset composed of 210 samples referring to three wheat varieties, 70 elements for each, described by 7 geometric features [29]. Considering the combination of the kernel data transformation (mapping) and the EPCA as an efficient preprocessing for input data, the Figure 4(a) illustrates the reduced number of features in the high-dimensional space (210 features). For the sake of clearness, we limit our plot to 25 features because of the decreasing evolution of the entropy. Therefore, the seven maintained nonlinear features interpret the more relevant ones in term of information content for our clustering algorithm. In the Figure 4(b) and 4(c) the displayed results are given based on the same magnitudes (the density and the distance). For the plot in Figure 4(b) it is a density-distance plot whereas the plot in Figure 4(c), it reveals the sorted product of density and distance. This quantity is by its definition large and begins to grow between cluster centroids progressively afterwards it becomes tight between the rest of points. Thus, the same result is given by both techniques, it is clearly seen that three centroids were determined from data, which declares in advance the existence of three clusters. The obtained results are confirmed thanks to the assumption that each center is distinguished by its high local density and relatively large distance away from the other data points with higher density. Hence, our algorithm is apt to extract centers from data points as a first step. To prove its efficiency, we opt to investigate the cluster validity index in Figure 4(d) and the Elbow method in Figure 4(e).

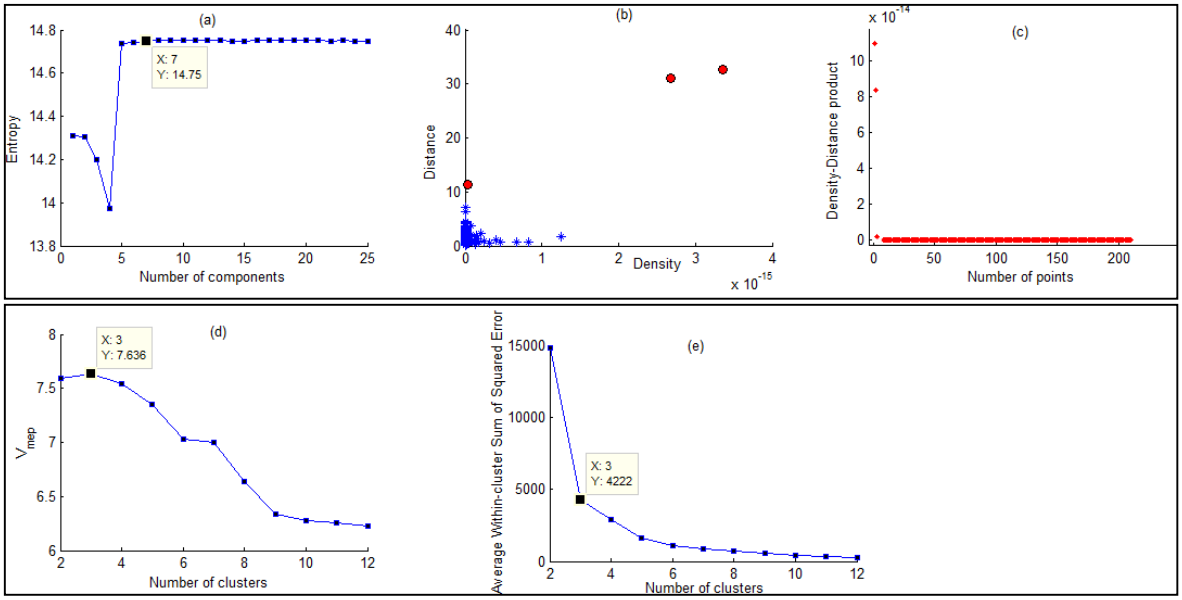

Figure 3. Results obtained on Seeds dataset (a) result of the optimal components number using the KEPCA, (b) decision graph for centers selection of the density distance plot, (c) the product of the density and distance plotted in decreasing order, (d) number of clusters validation given by V_{MEP} index and (e) the Elbow method for Seeds dataset

The classification rate given by our clustering algorithm considering the KEPCA data transformation is equal to 87.62% using VKE, whereas by using the Parzen estimator we got 89.52%, and considering EPCA data reduction with VKE we obtain 87.62% and the k-medoids and the k-means, results are equal to 84%. As a comparison, we conclude that our algorithm has better result than its counterpart algorithms of clustering.

3.4. Trajectories database

The LCPC (Laboratoire Central des Ponts et Chanssées) makes it possible to provide a database of experimental trajectories by measuring the parameters of trajectory in the bend during the vehicle passage at a discrete intervals of time. To achieve the purpose, they utilize a data acquisition system incorporated to a test vehicle that can restore the positions, speeds and accelerations. The survey was conducted with volunteer drivers, where each driver has to go through different trajectories. Hence, the database of physical trajectories was founded [31]. Therefore, in the present study, we are motivated to investigate the trajectories database,

which consists of 232 trajectories with 6 variables (longitudinal position, lateral position, longitudinal speed, lateral speed, longitudinal acceleration and lateral acceleration). Thus, we consider our approach for data transformation combining the kernel function the EPCA. The Figure 5 presents results obtained on trajectories dataset, Figure 5(a) demonstrates the reduced features in the high-dimensional space (232 features). We limit our plot to only 25 features on the Figure 5(a) for the sake of clearness since the entropy evolves in decreasing order. Hence, the three maintained nonlinear features are selected to have the highest value of entropy, which means they interpret the more relevant ones for our clustering algorithm in term of information content. Afterwards, on the Figures 5(b) and 5(c) both charts are given based on the pair density-distance measurement. The first plot on the Figure 5(b) presents a density-distance plot, whereas the Figure 5(c) illustrates the sorted quantity of the density-distance product. Hence, identical result was given by both techniques, four centroids have been determined. The obtained results are justified thanks to the assumption that defines every center to be distinguished from the other data points by its higher local density and relatively large distance. Thus, our algorithm is capable of differentiating between centers and the other points. To demonstrate our clustering algorithm performance, we investigate the cluster validity index in Figure 5(d) and the Elbow method in Figure 5(e). Both criteria perform same result, which confirms the existence of four clusters for the behavior of drivers. On Figure 5(f), we displayed the samples (trajectories) assignation to their appropriate clusters based on the centroids selection result where the remaining points are assigned to the closest centroid based on the measurement of the Euclidean distance. The given result is displayed on three-dimensional plot since the reduced dimension after performing KEPCA was deduced to be a three-feature Figure 5(a). The four clusters are discovered and clearly distinguished from one to another and represented by different colors and shapes. Each cluster represents a driver's behavior. The first class C1, corresponds to the family of the slowest trajectories of calmed driving. The second class C2, corresponds to the family of the slowest trajectories of sporting driving. The third class C3, represents the family of the fastest trajectories of calmed driving. The fourth class C4, correlates with the family of the fastest trajectories of sporting driving.

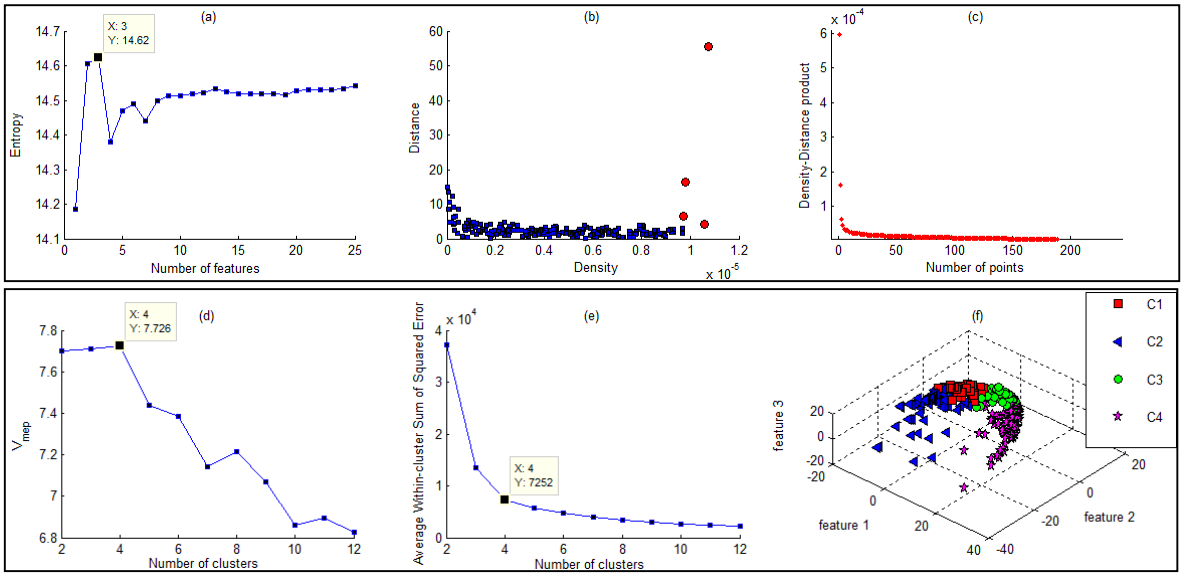

Figure 4. Results obtained on trajectories dataset (a) result of the optimal components number using the KEPCA, (b) decision graph for centers selection of the density distance plot, (c) the product of the density and distance plotted in decreasing order, (d) number of clusters validation given by V_{MEP} index and (e) the Elbow method, (f) Three-dimensional plot performed on mapped trajectories data in the reduced space of three features, samples are colored according to the cluster to which they are assigned with different color and shape

As a comparison to other works related to same trajectories dataset as in [31], same number of clusters and of driver's behavior were detected using our unsupervised algorithm based on kernel data transformation and variable kernel estimator for density estimation that is critical for clustering. The KEPCA has shown its potential over different dataset from artificial and real world database. Among all of them, the proposed KEPCA present an advantage of extracting only valuable information and mapping input data into high space, where the non-linearity can be omitted easily. The reduced number of features in the high space

improves the results. Firstly, in the PDF estimation using the reduced dimension where most of the entropy information is compacted and the selection of the kernel parameter become more robust. Secondly, in the clustering task as it is shown on the Table 1, the experiment results reveal the improvement of the performance of the algorithm using KEPCA over its counterpart EPCA. Furthermore, the use of VKE, often makes KEPCA more efficient than the use of Parzen estimator, which can be explained with the fact that the variable kernel estimator adapts the amount of smoothing to the local density data due to the adaptive scale that can vary from one data point to another.

Table 1. Comparison of the different clustering algorithms over artificial and real world data.

Dataset	Without data transformation	EPCA-VKE	KEPCA-Parzen	KEPCA-VKE	k-means
Artificial	96.83	82.33	96.67	97.17	96.83
Iris	-	-	88	89.33	89.33
Flame	84.17	80	85	85	85
heart	62.59	65.19	79.26	82.59	59.26
Trajectory	-	-	-	4 clusters	4 clusters
Seed	91.43	89.52	89.52	87.62	89.52

4. CONCLUSION

In the present paper, we propose an efficient data transformation for the existed EPCA method to extract the optimal features. Whereas the EPCA gives the optimal entropic component through maximizing the average of the information (the inertia) provided by the elements, the KEPCA indeed makes a mapping for data before considering EPCA. Therefore, the core element of the kernel used in KEPCA is to map implicitly the input data into a high-dimensional feature space, where the nonlinear patterns become linear and the separation of the elements becomes easier. We have revealed the ability of KEPCA to retain more information in the high space in both PDF estimation and clustering over synthetic and real world dataset examples. Results show the performance query of the KEPCA over its counterpart EPCA data transformation. Besides, the use of VKE estimator has proven its efficiency in density estimation, which has critical effect on the clustering algorithm.

REFERENCES

[1] D. Napoleon and S. Pavalakodi, "A New Method for Dimensionality Reduction Using KMeans Clustering Algorithm for High Dimensional Data Set," *International Journal of Computer Applications*, vol. 13, no. 7, pp. 41–46, Jan. 2011.

[2] R. H and A. T, "Feature Extraction of Chest X-ray Images and Analysis using PCA and kPCA," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 8, no. 5, pp. 3392–3398, Oct. 2018, doi: 10.11591/ijece.v8i5.pp3392-3398.

[3] A. K. Nikhath and K. Subrahmanyam, "Feature selection, optimization and clustering strategies of text documents," *International Journal of Electrical & Computer Engineering (IJECE)*, vol. 9, no. 2, pp. 1313–1320, Apr. 2019, doi: http://doi.org/10.11591/ijece.v9i2.pp1313-1320.

[4] J. MacQueen, "Some methods for classification and analysis of multivariate observations," *presented at the Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1967.

[5] A. K. Jain, "Data clustering: 50 years beyond K-means," *Pattern Recognit. Lett.*, vol. 31, no. 8, pp. 651–666, Jun. 2010, doi: https://doi.org/10.1016/j.patrec.2009.09.011.

[6] L. Kaufman and P. J. Rousseeuw, "Finding groups in data: an introduction to cluster analysis," *Hoboken, N.J: Wiley*, 2005.

[7] U. Ozertem, et al., "Mean shift spectral clustering," *Pattern Recognit.*, vol. 41, no. 6, pp. 1924–1938, Jun. 2008, doi: https://doi.org/10.1016/j.patcog.2007.09.009.

[8] Y. Weiss, "Segmentation using eigenvectors: a unifying view," *Proceedings of the Seventh IEEE International Conference on Computer Vision, Kerkyra, Greece*, vol. 2, 1999, pp. 975–982, doi: 10.1109/ICCV.1999.790354.

[9] E. mehdi Cherrat, Rachid Alaoui, Hassane Bouzahir., "Improving of Fingerprint Segmentation Images Based on K-MEANS and DBSCAN Clustering," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 9, no. 4, pp. 2425–2432, Aug. 2019, doi: 10.11591/ijece.v9i4.pp2425-2432.

[10] A. Rodriguez and A. Laio, "Clustering by fast search and find of density peaks," *Science*, vol. 344, no. 6191, pp. 1492–1496, Jun. 2014, doi: 10.1126/science.1242072.

[11] X.-F. Wang and Y. Xu, "Fast clustering using adaptive density peak detection," *Stat. Methods Med. Res.*, vol. 26, no. 6, pp. 2800–2811, Dec. 2017, doi: https://doi.org/10.1177%2F0962280215609948.

[12] L. E. Fattahi, et al., "Clustering based on density estimation using variable kernel and maximum entropy principle," *Intelligent Systems and Computer Vision (ISCV)*, 2017, pp. 1–7.

[13] K. Slaoui, "Optimisation des méthodes d'analyses factorielles, discriminante, et de classification des données statistiques et dynamiques par le principe de maximum d'entropie," *Doctoral thesis, University of Sidi Mohamed Ben Abdellah, Fez,* 2000.
[14] I. T. Jolliffe and J. Cadima, "Principal component analysis: a review and recent developments," *Philos. Trans. R. Soc. Math. Phys. Eng. Sci.,* vol. 374, no. 2065, 2016, doi: https://doi.org/10.1098/rsta.2015.0202.
[15] S. Damavandinejadmonfared and V. Varadharajan, "A New Extension of Kernel Principal Component Analysis for Finger Vein Authentication," *Comput. Sci.,* vol. 159, p. 5, 2015.
[16] E. T. Jaynes, "On the rationale of maximum-entropy methods," *Proc. IEEE,* vol. 70, no. 9, pp. 939–952, Sep. 1982, doi: 10.1109/PROC.1982.12425.
[17] O. Ammor, et al., "Optimal Fuzzy Clustering in Overlapping Clusters," *The International Arab Journal of Information Technology,* vol. 5, no. 4, p. 7, 2008.
[18] T. F. Abidin, et al., "Singular Value Decomposition for dimensionality reduction in unsupervised text learning problems," *2nd International Conference on Education Technology and Computer, Shanghai, China, 2010, pp. V4-422-V4-426,* doi: 10.1109/ICETC.2010.5529649.
[19] Z. Rustam and S. Hartini, "New feature selection based on kernel," *Bull. Electr. Eng. Inform.,* vol. 9, no. 4, pp. 1569-1577–1577, Aug. 2020, doi: https://doi.org/10.11591/eei.v9i4.1959.
[20] O. A. Adegbola, et al., "A principal component analysis-based feature dimensionality reduction scheme for content-based image retrieval system," *TELKOMNIKA Telecommun. Comput. Electron. Control,* vol. 18, no. 4, pp. 1892–1896, Aug. 2020, doi: 10.12928/TELKOMNIKA.v18i4.11176.
[21] K. M. Rao, et al., "Dimensionality reduction and hierarchical clustering in framework for hyperspectral image segmentation," *Bull. Electr. Eng. Inform.,* vol. 8, no. 3, pp. 1081-1087–1087, Sep. 2019, doi: http://dx.doi.org/10.11591/eei.v8i3.1451.
[22] C. E. Shannon, "A Mathematical Theory of Communication," *ACM SIGMOBILE mobile computing and communications review,* vol. 5, no. 1, pp. 3-55, 2001.
[23] L. E. Fattahi and E. H. Sbai, "Kernel entropy principal component analysis using Parzen estimator," *Int. Conf. Intell. Syst. Comput. Vis. ISCV,* pp. 1–8, 2018.
[24] B. W. Silverman, "Density Estimation for Statistics and Data Analysis," Boca Raton: Chapman and Hall/CRC, 1986.
[25] E. Parzen, "On Estimation of a Probability Density Function and Mode," *Ann. Math. Stat.,* vol. 33, no. 3, pp. 1065–1076, Sep. 1962.
[26] M. Rosenblatt, "Remarks on Some Nonparametric Estimates of a Density Function," *Ann. Math. Stat.,* vol. 27, no. 3, pp. 832–837, Sep. 1956.
[27] "UCI Machine Learning Repository." [Online]. Available: https://archive.ics.uci.edu/ml/index.php.
[28] D. J. Ketchen and C. L. Shook, "The Application of Cluster Analysis in Strategic Management Research: An Analysis and Critique," *Strateg. Manag. J.,* vol. 17, no. 6, pp. 441–458, 1996.
[29] M.Charytanowicz, et al.,"A Complete Gradient Clustering Algorithm for Features Analysis of X-ray Images," *Information technologies in biomedicine. Springer, Berlin, Heidelberg*, pp. 15-24, 2010..
[30] A. Koita, Dimitri Daucher, and Michel Fogli.., "Multidimensional risk assessment for vehicle trajectories by using copulas," *ICOSSAR, France,* 2013.
[31] A. Boubezoul, et al., "Vehicle trajectories classification using Support Vectors Machines for failure trajectory prediction," 2009 International Conference on Advances in Computational Tools for Engineering Applications., pp. 486–491, 2009, doi: 10.1109/ACTEA.2009.5227873.

BIOGRAPHIES OF AUTHORS

L. El Fattahi, she got a master degree on "signals, systems and informatics" from Faculty of Sciences, Dhar El Mehraz, University Sidi Mohamed Ben Abdallah, Fez, Morocco, in 2015. She has been working on his Phd thesis at the University of Moulay Ismaïl, Meknes, Morocco, with her research team "Control Systems and Information Processing (CSIP)" at High School of Technology of Meknes. Her scientific interests include data analysis, artificial intelligence, cluster analysis, image processing and machine learning.

S. El Hassan, received the Doctorate degree (Doctorat de 3ème cycle) from the University of Sidi Mohammed Ben Abdellah, Fez, Morocco in 1992, and the Doctorate of sciences (Doctorat d'Etat) from the university of Moulay Ismaïl, Meknes, Morocco in 2001. Since 1993, he is a Professor in automatics and informatics at High School of Technology of Meknes. His main research interest are cluster analysis, pattern recognition, machine learning and computer vision.