An adaptive algorithm for clustering cumulative probability distribution functions using the Kolmogorov-Smirnov two-sample test

Llanos Mora-López[a,*], Juan Mora[b]

[a] *Departamento de Lenguajes y Ciencias de la Computación, ETSI Informática Universidad de Málaga. Campus de Teatinos. 29071 Málaga. Spain*
[b] *Departamento de Fundamentos de Análisis Económico. Universidad de Alicante*

Abstract

This paper proposes an adaptive algorithm for clustering cumulative probability distribution functions (c.p.d.f.) of a continuous random variable, observed in different populations, into the minimum homogeneous clusters, making no parametric assumptions about the c.p.d.f.'s. The distance function for clustering c.p.d.f.'s that is proposed is based on the Kolmogorov-Smirnov two sample statistic. This test is able to detect differences in position, dispersion or shape of the c.p.d.f.'s. In our context, this statistic allows us to cluster the recorded data with a homogeneity criterion based on the whole distribution of each data set, and to decide whether it is necessary to add more clusters or not. In this sense, the proposed algorithm is adaptive as it automatically increases the number of clusters only as necessary; therefore, there is no need to fix in advance the number of clusters. The output of the algorithm are the common c.p.d.f. of all observed data in the cluster (the centroid) and, for each cluster, the Kolmogorov-Smirnov statistic between the centroid and the most distant c.p.d.f. The proposed algorithm has been used for a large data set of solar global irradiation spectra distributions. The results obtained enable to reduce all the information of more than 270000 c.p.d.f.'s in only 6 different clusters that correspond to 6 different c.p.d.f.'s.

Keywords: adaptive clustering, cumulative probability distribution

*Corresponding author
Email addresses: llanos@lcc.uma.es (Llanos Mora-López), juan@ua.es (Juan Mora)

functions, Kolmogorov-Smirnov two-sample test

1. Introduction

It is increasingly common to have a huge amount of data in many research fields, due to improved storage and easy access of the new computer systems. The challenges now facing researchers relate to the extraction of useful knowledge from all this stored information. In some systems, it is sufficient to use simple statistic summary of data such as mean and standard deviation of data. However, in many other systems, these simple statistics are not enough, and it is necessary to keep other statisitcal information, such as the distribution of the possible values that characterize the system. The empirical cumulative probability distribution function (c.p.d.f.) is a good tool to preserve inherent information such as variability and distribution of values. One of the problems with using this type of functions, when working with large amounts of data, is to determine how many different c.p.d.f.'s are necessary to keep all possible situations of the variable.

The analysis of empirical c.p.d.f.'s are useful in several domains, for instance: for storing data on sales for each customer, (Sakurai et al., 2008) and (Applegate et al., 2011); for analyzing images, (Spellman et al., 2005); for clustering images (Dontg et al., 2006) and (Lin et al., 2014); and for characterizing some meteorological parameters, (Mora et al., 2005), (Mora and Mora-López, 2010) and (Vrac et al., 2011).

Methods exist to decide whether or not the c.p.d.f.'s of two or more data sets are equal (homogeneously equal). However, when many empirical c.p.d.f.'s are obtained for the same variable (with data recorded from different populations), it would be useful to determine how many different c.p.d.f.'s really exist for that variable. That is, it may be interesting, or even necessary, to group all the available empirical c.p.d.f.'s in fewer clusters so that all c.p.d.f.'s in each cluster can be considered equal.

A model-based approach can be used to address the problem of clustering data according to (Montanari and Calo, 2013) in which a wavelet-based representation for the elements in the space is used and the clustering is accomplished by using mixture models for hyper-spherical data. Several different data mining techniques can be also used to face this problem. The goal of clustering data is to partition a data set into homogeneous clusters, see for instance the overviews of clustering in the literature (Ruspini, 1969),

(Hartigan, 1975) and (Jain and Dubes, 1988). Many different research areas have used clustering techniques such as text mining, statistical learning and pattern recognition (Jain et al., 1999), (Duda et al., 2001), (Hastie et al., 2001). Recently, different clustering methods and optimizations have been used to solve clustering tasks such as in (Lin et al., 2014) for image retrieval; in (Zhao et al., 2014) for image segmentation; in (Portela et al., 2014) for brain image fragmentation; in (Jun et al., 2014) to cluster documents.

Clustering techniques are based on using some type of distance functions, such as quadratic distance of Mahalanobis, Hausdorff distance or the Minkowsky metric, Jain and Dubes (1988). The Euclidean distance is one of the most common used among the different Minkowski distance metrics. It has been previously used to cluster cumulative probability distribution functions of solar spectral irradiance curves, (Moreno-Saéz and Mora-López, 2014). In that paper the authors work with a large amount of solar radiation spectra and analyze how many different spectra there are using the c.p.d.f. of each spectrum as a multidimensional variable representing the spectrum; using the k-means algorithm these curves are grouped into several distinct clusters.

Instead of using these metrics here we propose a more suitable metric for analyzing, comparing and clustering c.p.d.f.'s already used in Statistics that is based on the use of the Kolmogorov-Smirnov two sample statistic. Moreover, we propose the use of an adaptive learning algorithm for clustering all the observed c.p.d.f.'s that extensively uses the results of K-S test. The problem we address in this paper is to find a clustering method for c.p.d.f.'s of a continuous random variable into the minimum homogeneous clusters assuming that each set of observations is randomly drawn for an unknown distribution.

This paper is organized as follows. The c.p.d.f., the measure proposed to compare c.p.d.f.'s and the k-means clustering technique are described in the second section. In the third section, the proposed methodology and the adaptive algorithm for clustering data are described. The fourth section describes data used for checking the proposed methodology. In the fifth section, the results obtained when the proposed methodology is used for actual solar global irradiance spectra data are presented. Finally, the conclusions of the work are summarized in the sixth section.

2. Materials and methods

2.1. Comparing cumulative probability distribution functions

The cumulative probability distribution function (c.p.d.f.) of a random variable X, $F_X(\cdot)$, is defined as

$$F_X(t) = \Pr(X \leq t) \tag{1}$$

for any real number t. Given observations $\{X_i\}_{i=1}^n$ of the random variable X, the empirical c.p.d.f. is defined as $\hat{F}_X(t) \equiv n^{-1} \sum_{i=1}^n \mathbf{I}(X_i \leq t)$ for any real number t, where $\mathbf{I}(A)$ is the indicator function of event A, which takes the value 1 if A is true or 0 otherwise. If the observations are independent and identically distributed (i.i.d.), with the same distribution as X, it is well-known that the empirical c.p.d.f. is an appropriate estimate of the c.p.d.f. of X.

Assume that we are given i.i.d. observations $\{X_i\}_{i=1}^n$ of the random variable X, and i.i.d. observations $\{Y_i\}_{i=1}^m$ of the random variable Y. Note that X and Y might denote the same phenomenon observed at two different populations. Suppose that we want to test the null hypothesis

$$\mathrm{H}_0 : F_X(\cdot) = F_Y(\cdot), \tag{2}$$

versus the general alternative hypothesis

$$\mathrm{H}_a : F_X(\cdot) \neq F_Y(\cdot), \tag{3}$$

making no parametric assumption about the shape of these c.p.d.f.'s. This is known as the "test of homogeneity between two samples". If $F_X(\cdot)$ and $F_Y(\cdot)$ are continuous and the sample sizes n and m are large enough, the test can be performed using the Kolmogorov-Smirnov two-sample statistic, which compares the empirical c.p.d.f.'s obtained with each sample and is defined as

$$D_{\hat{F}_X(t),\hat{F}_Y(t)} = D_{n,m} \equiv \left(\frac{nm}{n+m}\right)^{1/2} \sup_{t \in \mathbb{R}} \left|\hat{F}_X(t) - \hat{F}_Y(t)\right|. \tag{4}$$

The null hypothesis is rejected with significance level α if $D_{n,m} > t_\alpha$, where t_α is a critical value that only depends on α; e.g. if $\alpha = 0.05$, then $t_\alpha = 1.36$ (for details, see Rohatgi and Saleh (2000)). Note that the Kolmogorov-Smirnov two-sample test enables to compare the empirical c.p.d.f.'s of two different samples without assuming any underlying parametric model for the samples,

i.e. it is a nonparametric test. An initial benefit of this type of test is that it does not impose any previous restriction of the data. Also note that that this test is able to detect differences in position, dispersion or shape of the c.p.d.f.'s of the two samples.

2.2. Clustering methods

The aim of clustering is to partition a data set (distributions) into groups, in such a way that one observation is more similar to the others of its cluster than to observations in other clusters according to some objective function that defines similarity or dissimilarity among objects (Han et al., 2006). It is based on analyzing one or more attributes to identify a cluster of correlating results.

Several partitional and hierarchical heuristic clustering methods have been proposed for clustering a set of observations, according the classification proposed by (Jain et al., 1999). Hierarchical algorithms recursively find nested clusters. They are not usually suitable for large data sets as they have quadratic or higher complexity in the size of samples. In contrast, partitional algorithms have lower complexity as they find all the clusters simultaneously as a partition of the data without imposing hierarchical structure. Both approaches are based on distance or dissimilarity measures. Hierarchical clustering algorithms produce partitions based on a criterion for merging or splitting clusters based on similarity. Partitional clustering algorithms obtain the partition that optimizes (locally) a clustering criterion.

Among the different clustering methods k-means, (MacQueen, 1967), is the most widely partitional clustering algorithm used (Celebi et al., 2013), (Jain et al., 1999) and (Celebi et al., 2013). The main problems of this method are that it may converge to a local minimum and that it depends on the selection of the initial centroids, as it has been pointed out by several authors, Jain et al. (1999), (Krishnasamy et al., 2014). Another problem of the classic k-means algorithm is that it tends to produce clusters with relatively uniform sizes as it finds it difficult to deal with imbalanced data, (Wu et al., 2007). Another of the limitations of the k-means technique is that the number of clusters needs to be fixed in advance.

Many heuristic approaches have been proposed over the years in order to address these problems. For instance, (Fathian et al., 2007) propose a honey-regarding bee-mating optimization; (Chen and Ye, 2004) and (Cura, 2012) propose a particle swarm optimization based aprorach; (Hatamlou, 2013) proposes a black hole optimization algorithm; and (Krishnasamy et

al., 2014) propose a hybrid approach based on modified cohort intelligence and k-means.

On the other hand, distance measures independent of clustering methods may not take into account the degradation of clustering performance, as has been pointed out by (Wu et al., 2009). They had shown that a data distribution view is of great use for selecting the right measures for clustering. The data for which we propose the adaptive clustering algorithm are cumulative probability distribution functions, using parameters related to these functions, such as the statistics proposed in this paper, the distance measure seems reasonable.

The k-means partition minimizes the sum, over all of the clusters, of the within-cluster sums of point-to-cluster-centroid distances. The squared Euclidean distance from the sample to its cluster was used to measure the similarity between each observation and the centroid of each cluster when using c.p.d.f. data. According to (Moreno-Saéz and Mora-López, 2014), algorithm 1 was used for clustering c.p.d.f.'s.

Input : K (the number of clusters) and the Training set ($\{x^{(1)}, x^{(2)}, \ldots, x^{(m)}\}$, where $x^{(i)} \in \mathbb{R}^n$ corresponds to $F(\lambda_i)$)

Randomly initialize K cluster centroids $\mu_1, \mu_2, \ldots, \mu_K \in \mathbb{R}^n$;

repeat

  for $i \leftarrow 1$ **to** m **do**

    $c^{(i)}$ = index j of the cluster centroid μ_j closest to $x^{(i)}$, $D_{x^{(i)},\mu_j} = \min\{D_{x^{(i)},\mu_k}\}, k = 1 \cdots K$ (using Eq. 4)

  end

  for $i \leftarrow 1$ **to** K **do**

    μ_i = the average of the points assigned to cluster i (this is the new centroid of the cluster);

  end

until *Assigned indices* $c^{(i)}$ *do not change*;

Output: Cluster for each sample (K clusters) and K centroids (c.p.d.f).

Algorithm 1: k-means algorithm.

3. Proposed methodology

3.1. Proposed adaptive learning algorithm for clustering c.p.d.f.'s

The problem of finding the best partition of a set of observations is np-hard problem. The problem can be formulated as follows. Given a set of empirical c.p.d.f.'s, $(y_1, y_2, \ldots, y_p)$, each of them obtained with n_i observations, i.e.:

$$y_i = F_{x_{(i)}}(t) = n_i^{-1} \sum_{j=1}^{n_i} \mathbf{I}(x_{(i)j} \leq t) \tag{5}$$

the objective is to partition the p c.p.d.f.'s into k sets $\{S_1, S_2, \ldots, S_k\} = S$ where k is the minimum possible number of sets that satisfies:

$$\begin{array}{l} \forall S_i \in S,\ i = 1 \ldots k,\ \forall pair(y_j^{(i)}, y_l^{(i)}) \in S_i, \\ D_{j,l} \equiv \left(\frac{n_j n_l}{n_j + n_l}\right)^{1/2} \sup_{t \in \mathbb{R}} \left| F_{x_{(j)}}(t) - F_{x_{(l)}}(t) \right| < t_\alpha. \end{array} \tag{6}$$

and n_j and n_l are the sizes of c.p.d.f.'s $y_j^{(i)}$ and $y_l^{(i)}$ respectively, both c.p.d.f.'s belonging to cluster i. It is also required that $k \leq k_{max}$, where k_{max} is the maximum number of clusters that could generate the algorithm. If the number of clusters is greater than k_{max}, the proposed algorithm obtains the maximum distance observed for each cluster (using the Kolmogorov-Smirnov two sample test). With this value it is possible to know the significance level of the hypothesis of homogeneity among all the c.p.d.f.'s in each cluster.

To achieve this objective, an adaptive clustering algorithm is proposed. This algorithm starts with all c.p.d.f.'s in one cluster and randomly selects one c.p.d.f. as the centroid of the cluster. Using this centroid, the empirical Kolmogorov-Smirnov statistic is estimated for this centroid and each one of the c.p.d.f.'s in the cluster. The following process is then repeated until all the estimated statistics are lower that the theoretical or the maximum number of specified cluster is reached:

- For each obtained cluster, obtain the maximum empirical Kolmogorov-Smirnov statistic estimated using the centroid of the cluster and each one of the c.p.d.f.'s of the cluster. If these maximum (one for each cluster) values are lower that the theoretical statistic, then stop (for an fixed significance level).

- Use the c.p.d.f. that corresponds to the maximum value among maximum values of all clusters as the centroid of a new cluster if the number of clusters is lower that the number of maximum specified clusters, otherwise stop.

- Assign each c.p.d.f. to the cluster for which the value of the Kolmogorov-Smirnov statistic estimated with the c.p.d.f. and the centroid of the cluster is lower.

Therefore, the process terminates either because all empirical statistics are less than the theoretical statistic or because the maximum number of specified clusters is reached. In the first case, it is possible to ensure that all analyzed c.p.d.f.'s can be represented only using the number of clusters created. In the latter case there may be two types of clusters:

- Those in which the empirical maximum distance observed (empirical statistic) is less than the theoretical statistic. All the c.p.d.f.'s are equal in these clusters.

- Those in which the empirical maximum observed distance is greater than the theoretical. In this case, the empirical statistic can be used to obtain the level of significance in the assumption of homogeneity in the cluster.

The output of the algorithm are the common c.p.d.f. of all observed data in the cluster (the centroid) and, for each cluster, the Kolmogorov-Smirnov statistic between the centroid and the most distant c.p.d.f

The proposed method is shown in algorithm 2.

Input : k_{max} (the maximum number of clusters) and the Training set $(\{y_1, y_2, \ldots, y_p\}$, where $y_j \in \mathbb{R}^n$ corresponds to $F_{x_{(j)}}(t))$

Initialize number of clusters: $k \leftarrow 1$

Randomly select one y_j (c.p.d.f.) and assign it to the centroid of cluster 1, $cpdf^{(1)} \leftarrow y_j \in \mathbb{R}^n$;

repeat

  for $j \leftarrow 1$ **to** p **do**

    c_j = index of the cluster centroid i for which $D_j = \min\{D_{i,j}\}\ i = 1 \ldots k$ according to Eq. 4; c.p.d.f. j is assigned to cluster i

  end

  for $i \leftarrow 1$ **to** k **do**

    $D^{(i)}_{max} \leftarrow \max\{D^{(i)}_j\}\ j = 1 \ldots l$ being l the number of c.p.d.f.'s in cluster i $c^{(i)}_{max} \leftarrow m \mid D_m \equiv D^{(i)}_{max}$

  end

  $D_k = \max\{D^{(i)}_{max}\}\ i = 1 \ldots k$,

  if $(D_k > t_\alpha) \wedge (k < k_{max})$ **then**

    $k \leftarrow k + 1$

    $cpdf^{(k)} = y_j$ being j the index of c.p.d.f. corresponding to D_{k-1}

  end

until $(D_k < t_\alpha) \vee (k > k_{max})$;

Output: Minimum number of clusters for c.p.d.f's and significance level of homogeneity hypothesis for each cluster.

Algorithm 2: Proposed adaptive algorithm.

The proposed adaptive selection of centroids is based on a similar methodology to the one proposed in the Maximim method, (Gonzalez, 1985) and (Katsavounidis et al., 1994), and in the k-means++ method, (Arthur and Vassilvitskii, 2007). In both approaches, the first center is also randomly selected and the following centroids are selected using the greatest minimum-distance in Maxmin method and a variant that chooses centers at random but weighs the data points according to their squared distance from the already chosen closest center in k-means++. However both methods always divide the samples into the previously specified number of clusters. In our proposal, the number of clusters can be fixed in advance but it is also possible to decide (and minimize) automatically the number of clusters using the significance level of the Kolmogorov-Smirnov two sample test used for

estimating the centroids.

The proposed algorithm is $O(kn)$, where n is the number of observations and k the number of clusters according (Gonzalez, 1985). Moreover, this algorithm guarantees solutions with an objective function value within twice the optimal solution value.

4. Data description

The proposed methodology was checked for solar spectral irradiance measurements recorded at the Photovoltaic Systems Laboratory of the University of Malaga. Each observation is composed of 920 variables (920-dimensional space). A Grating Spectroradiometer prepared for continuous outdoor exposure was to record them. It shortens the measurement to the range of 10 msec to 5sec. The geographical coordinates of the Laboratory are latitude 36.7^o N and longitude 4.5^o W, height 50 m. Measurements were collected from November 2010 to May 2012. Spectra were obtained with a spectral resolution of below 8 nm at a wavelength interval of 0.75 nm. For this study, we used the irradiance values that correspond to spectra whose wavelengths range from 350 to 1050 nm. A total of 920 values were used for each spectrum, and a total of 282,318 spectrum (samples) were used.

5. Results and discussion

We checked whether two solar spectral irradiance distributions are the same to decide how many different solar spectral irradiance distributions are in the recorded data using the proposed adaptive clustering method. Formally, we obtained the following:

$$\{X_{\lambda_i}\}_{\lambda_1=350}^{\lambda_n=1050} \quad \texttt{and} \quad \{Y_{\lambda_i}\}_{\lambda_1=350}^{\lambda_n=1050}$$

which are the solar spectral irradiance values for the different wavelengths λ_i of two measurements. Denote:

$$\hat{f}_X(\lambda_j) \equiv \frac{X_{\lambda_j}}{E_t^{(X)}} \tag{7}$$

where $E_t^{(X)}$ is the total amount of energy received for all wavelength for spectra X, according to Ec.8:

$$E_t^{(X)} = \sum_{\lambda_i=350}^{1050} X_{\lambda_i} \tag{8}$$

The sample sizes for this experiment are 920 as it is the number of solar spectral irradiances measured (one for each wavelength recorded by the measurement equipment). Specifically, for a real number λ_j in the range $[350.0 - 1050.0]$ (which corresponds to one of the wavelengths measured), we define:

$$\hat{F}_X(\lambda_j) \equiv \sum_{\lambda_i=350}^{\lambda_j} \hat{f}_X(\lambda_i) \quad \texttt{and} \quad \hat{F}_Y(\lambda_j) \equiv \sum_{\lambda_i=350}^{\lambda_j} \hat{f}_Y(\lambda_i) \tag{9}$$

and for these functions we checked the hypothesis of Eq.2 by estimating the Kolmogorov-Smirnov statistic, Eq.4, in the proposed adaptive algorithm 2 (significance level=0.05).

First, we checked the performance of the adaptive algorithm. Table 1 shows the maximum distance estimated in each cluster for each iteration when the algorithm is executed. Initially, all the data are in one only cluster and the number of clusters is increased in 1 more cluster in each iteration. The algorithm stops when all the clusters meet the homogeneity test or when the number of clusters is equal to the maximum allowed clusters. In this case, the algorithm ends when there are 6 clusters.

Number of clusters	Max distance for homogeneity test					
	1	2	3	4	5	6
1	0.119					
2	0.078	0.077				
3	0.068	0.041	0.053			
4	0.052	0.041	0.053	0.038		
5	0.047	0.041	0.050	0.038	0.051	
6	0.038	0.039	0.038	0.038	0.038	0.036

Table 1: Maximum distances in each cluster using the proposed adaptive algorithm

The obtained results agree with previously reported results obtained for clustering c.p.d.f.'s of solar spectra irradiance data, see (Moreno-Saéz and Mora-López, 2014). The advantages of the proposed method are that it is

possible to ensure that all c.p.d.f.'s in each cluster fulfill the Kolmogorov-Smirnov two sample test and that the c.p.d.f.'s are clustered using the minimum number of different clusters without a value for the number of clusters being previously set.

Second, we checked the algorithm by executing it several times and analyzing the obtained results. Moreover, we executed the classical k-means algorithm also using the Kolmogorov-Smirnov two-sample test as metric. The number of times both algorithms were executed is 20. For each execution, we obtained the maximum distance among the c.p.d.f.'s in each cluster and the number of clusters in which the homogeneity test among the data is not met when the classical k-means and the proposed adaptive clustering algorithm are used. The maximum distance observed and the number of iterations in each execution for both algorithms are shown in Figure 1.

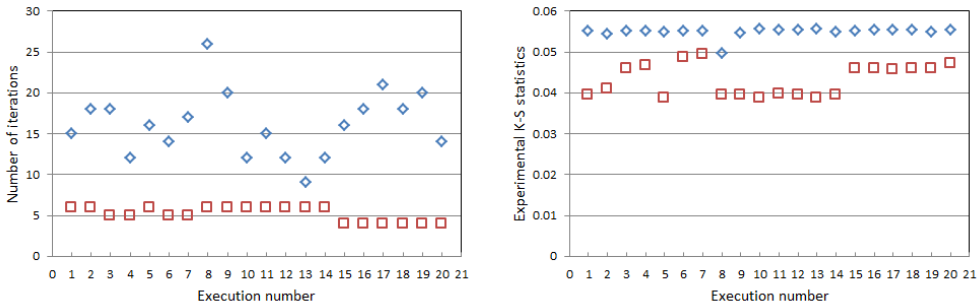

Figure 1: Number of iterations and maximum experimental Kolmogorov-Smirnov statistics in each execution of the adaptive proposed algorithm □ and the classical k-means algorithm ◇.

As can be observed, the proposed adaptive algorithm is able to cluster observations using fewer iterations than the classical k-means as only five or six iterations are necessary in most cases while the classic k-means algorithm required between 10 and 20 iterations. Moreover, the Kolvomorov-Smirnov experimental statistics (maximum distances) to the groups obtained with the proposed algorithm are in all cases lower than the critical values for significance levels smaller than 0.2 while those obtained with the classical algorithm are smaller than the critical values for significance levels smaller than 0.1 significance. This suggests that the proposed algorithm clusters better the observations taking into account the Kolmogorov-Smirnov two sample test.

We have also analyzed the size of the different clusters. Figure 2 shows the distribution of the samples in each cluster for two different executions of both algorithms.

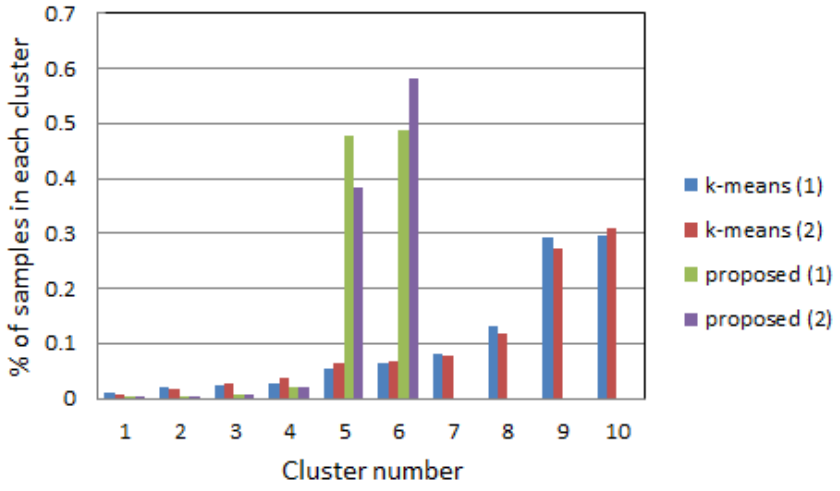

Figure 2: Percentage of samples in each cluster for two different executions of the classical k-means and the proposed adaptive algorithms.

As can be observed, the proposed adaptive algorithm generates fewer clusters (6 versus 10 generated by the classical k-means) and the generated clusters have very different sizes while the classical k-means tends to distribute the observations in clusters more homogeneous, as it has been pointed out by (Wu et al., 2007). The algorithm is valid for imbalanced data sets as it does not tend to produce clusters with similar sizes. Conversely, the proposed algorithm relaxes the constraints of homogeneity in all cluster, also in the case that the number of clusters is greater than the maximum prefixed. In this case, the algorithm obtains an approximation of the significance level of homogeneity in each cluster.

6. Conclusions

An adaptive algorithm for clustering cumulative probability distribution functions of a continuous random variable into the minimum homogeneous clusters is proposed. The main contributions of the work are the proposed use

of a new distance function for clustering c.p.d.f.'s and the automatic selection of the number of necessary clusters (limited by a prefixed maximum value).

The Kolmogorov-Smirnov two sample statistic is proposed as a distance function in the algorithm. This statistic is used to establish the homogeneity of observations in each cluster. This new distance function allows us to use significant statistic information in the observation-cluster process.

The number of clusters does not need to be fixed in advance as the adaptive algorithm is able to decide when it is necessary to add new clusters. Specifically, the decision to add a new cluster depends on the Kolmogorov-Smirnov two sample statistic of each cluster and the specified significance level for the homogeneity of clusters.

The proposed algorithm relaxes the homogeneity constraints in all clusters in the case that the number of clusters is greater than the maximum prefixed. In this case, the algorithm is capable of obtaining either the minimum number of clusters that meet the Kolmogorov-Smirnov two-sample test for a fixed significance level or an approximation of the significance level of homogeneity in each cluster in which the Kolmogovor-Smirnov test rejects the equality among the data in the cluster.

The algorithm is valid for imbalanced data sets as it does not tend to produce clusters with similar size but focuses on obtaining clusters in which the Kolmogorov-Smirnov statistics is minimum (it is the proposed distance measurement among c.p.d.f.'s).

The algorithm has been checked using actual data. The obtained results show that the algorithm is capable of clustering a large amount of cumulative probability functions in only 6 clusters, ensuring homogeneity in all cluster (significance level 0.05). Moreover, the algorithm achieves better distribution of observations in each cluster than the classical k-means algorithm.

Future research lines are, first, to analyze whether the algorithm is useful for other types of data that also use c.p.d.f.'s, such as image processing, clustering documents or brain image fragmentation. Furthermore, it would be of interest to improve the algorithm both to allow a fast estimation of Kolmogorov-Smirnov statistic by reusing the information estimated in previous iterations and to enable the detection of outliers.

Acknowledgments

This research has been partially supported by the Spanish Consejería de Economía, Innovación y Ciencia of the Junta de Andalucía under projects

TIC-6441 and P11-RNM7115, and the Spanish MEC under project ECO2011-29751.

Applegate, D., Dasu, T., Krishnan, S., Urbanek, S., 2011. Unsupervised clustering of multidimensional distributions using earth mover distance. In: ACM (Ed.), 17th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, pp. 636–644.

Arthur, D., Vassilvitskii, S., 2007. k-means++: The advantages of careful seeding. In: ACM-SIAM Symposium on Discrete Algorithms (SODA).

Celebi, M., A., K. H., Vela, P., 2013. A comparative study of efficient initialization methods for the k-means clustering algorithm. Expert Systems with Applications 40, 200–210.

Chen, C.Y., Ye, F. 2004. Particle swarm optimization algorithm and its application to clustering analysis. In IEEE international conference on networking, sensing and control, 2, 789-794.

Cura, T. 2012. A particle swarm optimization approach to clustering. Expert Systems with Applications, 39, 15821588.

Dong, L., Ogunbona, P., Li, W., Yu, G., Fan, L., Zheng, G. 2006. A fast algorithm for color image segmentation. In: Proceedings of International Conference on Image Processing, 685688.

Duda, R., Hart, P., Stork, D., 2001. Pattern classification. John Wiley & Sons.

Fathian, M., Amiri, B., Maroosi, A. 2007. Application of honey-bee mating optimization algorithm on clustering. Applied Mathematics and Computation, 190, 15021513.

Gonzalez, T., 1985. Clustering to minimize the maximum intercluster distance. Theoretical Computer Science 38, 293–306.

Han, J., Kamber, M., Pei, J., 2006. Data Mining: Concepts and Techniques, 2nd Edition. Morgan Kaufmann.

Hartigan, J., 1975. Clustering Algorithms. Wiley.

Hastie, T., Tibshirani, R., Friedman, J., 2001. The elements of statistical learning: Data mining, inference and prediction. Springer.

Hatamlou, A. 2013. Black hole: A new heuristic optimization approach for data clustering. Information Sciences, 222, 175184.

Jain, A. K., Dubes, R., 1988. Algorithms for Clustering. Englewood Cliffs, N.J.

Jain, A. K., Murty, M. N., Flynn, P. J., Sep. 1999. Data clustering: a review. ACM Computing Surveys 31 (3), 264–323.

Jun, S., Park, S.S., Jang, D.S., 2014. Document clustering method using dimension reduction and support vector clustering to overcome sparseness. Expert Systems with Applications, 41, 3204-3212.

Katsavounidis, J., Kuo, J., Zhang, Z., 1994. A new initialization technique for generalized lloyd iteration. IEEE Signal Processing Letters 1 (10), 144–146.

Krishnasamy, G., Kulkarni, A.J., Paramesran, R. 2014. A hybrid approach for data clustering based on modified cohort

Chuen-Horng Lin, Chun-Chieh Chen, Hsin-Lun Lee, Jan-Ray Liao. 2014. Fast K-means algorithm based on a level histogram for image retrieval. Expert Systems with Applications, 41 (7), 3276-3283.

MacQueen, J., 1967. Some methods for classification and analysis of multivariate observations. In: Proceedings of the 5th Berkeley symposium on mathematical statistics and probability. Statistics (Vol.1) University of California Press.

Montanari, A., Calo, D., 2013. Model-based clustering of probability density functions. Advances in Data Analysis and Classification 7 (3), 301–319.

Mora-López, L., Mora, J., Morales-Bueno, R., Sidrach-de-Cardona, M. 2005. Modeling time series of climatic parameters with probabilistic finite automata. Environmental Modelling & Software, 20 (6), 753760.

Mora, J., Mora-López, L., 2010. Comparing distributions with bootstrap techniques: An application to global solar radiation. Mathematics and Computers in Simulation 81, 811–819.

Moreno-Saéz, R., Mora-López, L., 2014. Modelling the distribution of solar spectral irradiance using data mining techniques. Environmental Modelling & Software 53, 163–172.

Moreno-Saéz, R., Sidrach-de Cardona, M., Mora-López, L., 2013. Data mining and statistical techniques for characterizing the performance of thin-film photovoltaic modules. Expert Systems with Applications 40, 7141–7150.

Portela, N.M., Cavalcanti, G.D.C., Ren T.I. 2014. Semi-supervised clustering for MR brain image segmentation. Expert Systems with Applications, 41, 1492-1497.

Rohatgi, V. K., Saleh, A. K. M. E., 2000. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., pp. 717–725.

Ruspini, E., 1969. A new approach to clustering. Information Control 15 (1), 22–32.

Sakurai, Y., Chong, R., Lei, L.and Faloutsos, C., 2008. Efficient distribution mining and classification. In: Proceedings of the 2008 SIAM international conference on data mining.

Spellman, E., Vemuri, B., Rao, M., 2005. Using the kl-center for efficient and accurate retrieval of distributions arising from texture images. IEEE Comput Soc Confer Comput V Pattern Recogn 1, 111–116.

Vrac, M., L., B., Diday, E., Chdin, A., 2011. Copula analysis of mixture models. Computational Statistics 27, 427–457.

Wu, J., Chen, J., Xiong, H., Xie, M., 2009. External validation measures for k-means clustering:a data distribution perspective. Expert Systems with Applications 36, 6050–6061.

Wu, J., Xiong, H., Chen, J., Zhou, W., 2007. A generalization of proximilty functions for k-means. In: Proceedings of the 2007 IEEE International Conference on Data Mining.

Zhao, F., Fan, J., Liu, H. 2014. Optimal-selection-based suppressed fuzzy c-means clustering algorithm with self-tuning non local spatial