Expert Systems with Applications 41 (2014) 3351–3366

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Dynamic clustering of histogram data based on adaptive squared Wasserstein distances

Antonio Irpino [a,*], Rosanna Verde [a], Francisco de A.T. De Carvalho [b]

[a] Dipartimento di Scienze Politiche "J. Monnet", Second University of Naples, 81100 Caserta, Italy
[b] Centro de Informatica – CIn/UFPE, Av. Prof. Luiz Freire, s/n, Ciadade Universitaria, CEP 50.740-540 Recife, PE, Brazil

ARTICLE INFO

Keywords:
Histogram data
Partitioning clustering method
Wasserstein distance
Adaptive distance
Symbolic data analysis

ABSTRACT

This paper presents a Dynamic Clustering Algorithm for histogram data with an automatic weighting step of the variables by using adaptive distances. The Dynamic Clustering Algorithm is a *k*-means-like algorithm for clustering a set of objects into a predefined number of classes. Histogram data are realizations of particular set-valued descriptors defined in the context of Symbolic Data Analysis. We propose to use the ℓ_2 Wasserstein distance for clustering histogram data and two novel adaptive distance based clustering schemes. The ℓ_2 Wasserstein distance allows to express the variability of a set of histograms in two components: the first related to the variability of their averages and the second to the variability of the histograms related to different size and shape. The weighting step aims to take into account global and local adaptive distances as well as two components of the variability of a set of histograms. To evaluate the clustering results, we extend some classic partition quality indexes when the proposed adaptive distances are used in the clustering criterion function. Examples on synthetic and real-world datasets corroborate the proposed clustering procedure.

1. Introduction

In many real experiences, data are grouped and summarized by histograms. For example, in the framework of image analysis, the characteristics of the images can be represented as histograms (even if they have to be considered as bar diagrams). Histogram descriptions are used for privacy preserving matters (for example, the cash flows of a bank account), as well as for the dissemination of official statistics, or when it is more relevant the aggregated information than the single observations. *Histogram data* formalization (in terms of descriptions of statistical units) were introduced in the context of *Symbolic Data Analysis* (SDA) by Bock and Diday (2000) as particular set-valued descriptions. In this framework, several techniques have been proposed for the statistics treatment of such new entities.

A classic tool for the exploration of a set of data is the cluster analysis, which aims to collect a set of objects in a number of homogeneous clusters according to the values they assume with respect to a set of observed variables. Clustering techniques may be divided into hierarchical and partitioning methods (Jain, 2010; Xu & Wunusch, 2005). Among the partitioning methods, Dynamic Clustering (DC) (Diday & Simon, 1976), a generalization of the *k-means* algorithm, showed some interesting properties in treating set-valued descriptions. The DC method (Diday & Simon, 1976) is a general partitioning algorithm of a set of objects in *K* clusters. It is a two step algorithm that minimizes a within homogeneity criterion and looks for the best representation of each cluster according to the homogeneity criterion. In DC, the choice of a suitable dissimilarity plays a central role for the definition of the allocation and of the representation phases. The *k-means* algorithm is a particular case of DC where the criterion function is expressed as the sum of the squared Euclidean distances of the objects with respect to the mean of the belonging cluster. According to the nature of data and the chosen dissimilarity function, DC is a more general schema of partition around a set of prototypes. In the case of the *k-means*, prototypes are the means of each cluster, while the DC can admit more general prototypes, like a sets of elements of the cluster, regression lines, factorial axes and so on.

A main issues in clustering analysis is to take into account the different contribution of each variable in the clustering process according to their variability. Conventional clustering algorithms do not take into account the relevance of the variables, i.e., these algorithms consider that all variables are equally important to the clustering process. However, in most applications some variables may be irrelevant and, among the relevant ones, some may be more or less relevant than others. Furthermore, the relevance of each variable to each cluster may be different, i.e., each cluster may have a different set of relevant variables. To face this problem,

* Corresponding author.
E-mail addresses: antonio.irpino@unina2.it, irpino@unina.it (A. Irpino), rosanna.verde@unina2.it (R. Verde), fact@cin.ufpe.br (Francisco de A.T. De Carvalho).

0957-4174/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2013.12.001

it is usual to standardize data in order to allow to each variable playing a comparable role in the analysis. However such strategy cannot take into account the importance of each variable in the clustering process. In order to tackle this issue Diday and Govaert (1997) proposed to integrate adaptive distances. The use of adaptive distances in the clustering algorithm is done introducing a weighting step in the optimization process. In this step a set of weights are obtained minimizing the total sum of squares criterion. Such weights are associated with each variable (for all the clusters or for each cluster) and represents a measure of the importance of a variable in the clustering process. More recent approaches to compute the relevance weight of a variable in the clustering process can be found in Ref. Frigui and Nasraoui (2004), Chan, Ching, Ng, and Huang (2004), Friedman and Meulman (2004), Huang, Ng, Rong, and Li (2005), Jing, Ng, and Huang (2007), Tsai and Chiu (2008), Deng, Choi, Chung, and Wang (2010), Ahmad and Dey (2011) and Chen, Ye, Xu, and Huang (2012). In the framework of SDA, De Carvalho and Lechevallier (2009a, 2009b), De Souza and De Carvalho (2007) and De Carvalho and De Souza (2010) proposed several adaptive distances (based on Hausdorff, City-Block and Euclidean distances) in Dynamic Clustering Algorithm of set-valued data.

Clustering methods are generally based on dissimilarity/similarity measures for comparing data. In the special field of image analysis Rubner, Tomasi, and Guibas (2000) introduced the Earth Mover's distance (EMD). It is worth to note that EMD between histograms of pixel intensities is equivalent to the Mallow's, or ℓ_2 Wasserstein distance (Rüshendorff, 2001; Villani, 2003) for probability distributions (Levina & Bickel, 2001; Mallows, 1972) when the histogram of pixel counts are normalized to one. However, the comparison of histogram data can be seen as a particular case of comparison of probability distribution functions of random variables. The first formulations of this distance for statistical purposes goes to the Gini's studies in 1918. To this aim, several distances for histograms have been presented in the literature (in the SDA framework a survey is available in Verde & Irpino (2008b)).

Several proposals have been presented in the SDA literature for clustering histogram data (see Irpino & Verde (2006), Verde & Irpino (2006, 2008a, 2008b)). More recently, Terada and Yadohisa (2010) proposed a *k*-means clustering method using empirical joint distributions. Vrac, Billard, Diday, and Chedin (2012) give a Dynamic Clustering Algorithm based on the use of copula analysis aiming to take into account the relationship between the histogram variables. Calo, Montanari, and Viroli (2014) presented a hierarchical mixture model that allows dimension reduction by assuming a generative factorial model for the observed histogram variables. Despite these recent contributions to clustering analysis of histogram data, none of them is able to compute automatically a relevance weight of each histogram variable during the clustering process.

The present paper present a Dynamic Clustering Algorithm based on the use of the ℓ_2 Wasserstein distances to compute the dissimilarity between histogram data. Thanks to two novel adaptive distance based clustering schemes, the proposed method is able to compute automatically the relevance weight of each histogram variable during the partitioning of the data set. Ref. De Carvalho and De Souza (2010) gives also a clustering algorithm with automatic weighting of the histogram variables. However, Ref. De Carvalho and De Souza (2010) uses an Euclidean distance between two sets of weights related to a particular pre-processing of the set-valued data. In the present paper, the ℓ_2 Wasserstein distance does not require pre-processing of the input histograms and it is not affected by different schemes of binning for the histograms. Further, using a particular decomposition of the ℓ_2 Wasserstein distance (Irpino & Romano, 2007) and considering the variability measure introduced in Verde and Irpino (2008b), it is possible to express the variability of a set of histograms in two parts: the first related to the variability of averages of the histograms and the second related to the variability due to the different sizes or shapes of the histograms. Thus, it is possible to consider the ℓ_2 Wasserstein distance as measure of diversity of two distributions according to two (additive) sources (or components) of variability.

In order to take advantage from this decomposition, we propose a global and a local approach for the definition of adaptive distances that take into account the two components of the variability of a set of histograms. In the global approach, we propose to associate two sets of weights to each variable and to each component. The two sets are globally estimated for all the clusters at once. In the local approach we consider also a different set of weights for each cluster.

Moreover, we prove the decomposition of the total inertia of a set of histogram data computed with the adaptive (squared) Wasserstein distances in within (intra-cluster) and between (inter-cluster) inertia. According to this result, we provide clustering interpretative tools based on the extension of the classic quality partition indexes (Celeux, Diday, Govaert, Lechevallier, & Ralambondrainy, 1989).

This paper is organized as follows: in Section 2, we introduce the definitions of histogram data and the Wasserstein distance between histograms. In Section 3, starting from the Dynamic Clustering Algorithm with non-adaptive distances, we propose two schemes where the adequacy criterion is based on adaptive squared Wasserstein distances. The first, we denote as *Globally Component-wise Adaptive Wassertein Distance* (GC-AWD), while the second, as *Cluster Dependent Component-wise Adaptive Wassertein Distance* (CDC-AWD). In Section 3.2, we introduce some tools for the interpretation of the clustering results. In Section 4, two applications are shown: one using synthetic data in order to illustrate the usefulness of the proposed methods based on the variability structure of the data; the other one, using a real dataset in order to demonstrate the application in a real situation and to show how to interpret the results of a classic clustering task on histogram data. Section 5 ends the paper with some conclusions and perspectives about the proposed clustering methods.

2. Histogram data and Wasserstein distance

Histogram is a suitable (in terms of computational resources) way for the representation of aggregate data or empirical distributions. SDA formalized histogram data as realizations of a histogram variable (a special case of modal-valued variable). In this case, the variable Y is a histogram-valued variable if to each observation i corresponds a probability or a frequency distribution described by a histogram (Bock & Diday, 2000).

Formally, let y_i a realization of Y such that $S(i) = [min(y); Max(y)] \subset \Re$ is the support, that is partitioned into a set of contiguous intervals (bins) $\{I_{1i}, \ldots, I_{hi}, \ldots, I_{H_i i}\}$ (where $I_{hi} = [a_{hi}; b_{hi})$ with $min(y) = a_{1i}$ and $Max(y) = b_{H_i i}$) and each I_{hi} is associated with a (non negative) weight π_{hi} that represents an empirical (or theoretical) relative frequency. In this paper, we denote with $f_i(y)$ the (empirical) density function associated with the description y_i and with $F_i(y)$ its cumulative distribution function. It is possible to define the description of the ith histogram for the variable Y as:

$$y_i = \left[(I_{1i}, \pi_{1i}), \ldots, (I_{ui}, \pi_{ui}), \ldots, (I_{H_i i}, \pi_{H_i i})\right]$$
$$\text{such that } \forall I_{ui} \in S(i), \quad \pi_{ui} = \int_{I_{ui}} f_i(y)dy \geqslant 0 \text{ and } \int_{S(i)} f_i(y)dy = 1. \tag{1}$$

In the following, we use y_i to denote the histogram associated with the ith unit when a single histogram variable is observed. If we obverse p variables, we denote with y_{ij} (where $i = 1, \ldots, n$ and