Dynamic clustering of histogram data based on adaptive squared Wasserstein distances

Antonio Irpino\textsuperscript{a,}\textsuperscript{*}, Rosanna Verde\textsuperscript{a}, Francisco de A.T. De Carvalho\textsuperscript{b}

\textsuperscript{a} Dipartimento di Scienze Politiche “J. Monnet”, Second University of Naples, 81100 Caserta, Italy
\textsuperscript{b} Centro de Informatica – CIn/UFPE, Av. Prof. Luiz Freire, s/n, Ciadidade Universitaria, CEP 50.740-540 Recife, PE, Brazil

**A R T I C L E   I N F O**

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

**A B S T R A C T**

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 \textit{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 $\ell_2$ Wasserstein distance for clustering histogram data and two novel adaptive distance based clustering schemes. The $\ell_2$ 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.

© 2013 Elsevier Ltd. All rights reserved.

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,
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 \( \ell_2 \) Wasserstein distance (Rüshendorf, 2001; Villani, 2003) for probability distributions (Levina & Bickel, 2001; Mallows, 1972) when probability distributions (Levina & Bickel, 2001; Mallows, 1972) when considering 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 \( S(i) = \{\min(y) \leq \cdot \leq \max(y)\} \subset \mathcal{R} \) is the support, that is partitioned into a set of contiguous intervals (bins) \( \{I_1, \ldots, I_{n_b}\} \) where \( I_{b_1} = [a_{b_1}, b_{b_1}) \) with \( \min(y) = a_1 \) and \( \max(y) = b_{n_b} \) and each \( b_{n_b} \) is associated with a (non negative) weight \( \pi_{n_b} \) 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 \( h \)-th histogram for the variable \( Y \) as:

\[
Y_i = \left( (I_{h_1}, \pi_{h_1}), \ldots, (I_{h_{n_h}}, \pi_{h_{n_h}}) \right)
\]

such that \( \forall I_{b_1} \in S(i), \quad \pi_{b_1} = \int_{I_{b_1}} f_i(y) \, dy \geq 0 \) and \( \int_{I_{b_1}} f_i(y) \, dy = 1 \).

(1)

In the following, we use \( y_i \) to denote the histogram associated with the \( i \)-th unit when a single histogram variable is observed. If we obverse \( p \) variables, we denote with \( y_{ij} \) (where \( i = 1, \ldots, n \) and
دریافت فوری
متن کامل مقاله

امکان دانلود نسخه تمام متن مقالات انگلیسی
امکان دانلود نسخه ترجمه شده مقالات
پذیرش سفارش ترجمه تخصصی
امکان جستجو در آرشیو جامعی از صدها موضوع و هزاران مقاله
امکان دانلود رایگان ۲ صفحه اول هر مقاله
امکان پرداخت اینترنتی با کلیه کارت های عضو شتاب
دانلود فوری مقاله پس از پرداخت آنلاین
پشتیبانی کامل خرید با بهره مندی از سیستم هوشمند رهگیری سفارشات