Applied Soft Computing 14 (2014) 210–228

Contents lists available at ScienceDirect

Applied Soft Computing

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

Evolving soft subspace clustering

Lin Zhu [a,b,c,*], Longbing Cao [b], Jie Yang [a], Jingsheng Lei [c]

[a] Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai, China
[b] Faculty of Engineering and IT, Advanced Analytics Institute, University of Technology, Sydney, Australia
[c] School of Computer Science and Technology, Shanghai University of Electric Power, Shanghai, China

ARTICLE INFO

Article history:
Received 31 August 2012
Received in revised form 21 January 2013
Accepted 3 March 2013
Available online 22 March 2013

Keywords:
Subspace clustering
Data stream clustering
Online clustering
Scalable clustering

ABSTRACT

A key challenge to most conventional clustering algorithms in handling many real world problems is that, data points in different clusters are often correlated with different subsets of features. To address this problem, subspace clustering has attracted increasing attention in recent years. In practical data mining applications, data points may arrive in continuous streams with chunks of samples being collected at different time points. In addition, huge amounts of data often cannot be kept in the main memory due to memory restriction. Accordingly, a range of evolving clustering algorithms has been proposed, however, traditional evolving clustering methods cannot be effectively applied to large-scale high dimensional data and data streams. In this study, we extend the online learning strategy and scalable clustering technique to soft subspace clustering to form evolving soft subspace clustering. We propose two online soft subspace clustering algorithms, OFWSC and OEWSC, and two streaming soft subspace clustering algorithms, SSSC_F and SSSC_E. The proposed evolving soft subspace clustering leverages on the effectiveness of online learning scheme and scalable clustering methods for streaming data by revealing the important local subspace characteristics of high dimensional data. Substantial experimental results on both artificial and real-world datasets demonstrate that our proposed methods are generally effective in evolving clustering and achieve superior performance over existing soft subspace clustering techniques.

1. Introduction

Clustering has long been a hot research topic in various disciplines as an important data processing technique. It has been widely utilized as a fundamental tool for data analysis and visualization in areas such as data mining and machine learning [1]. Clustering aims to categorize unlabeled input data vectors into different groups, called clusters, such that data points within a cluster are more similar to one another than they are to data points belonging to different clusters, i.e., by maximizing the intra-cluster similarity while minimizing the inter-class similarity [2,3]. However, a key challenge to most conventional clustering algorithms is that, in many real world problems, data points in different clusters are often correlated with different feature subsets. For example, given a subset of data points identified in a cluster, it is possible that these points exhibit the same characteristics as points from other clusters when a certain subset of dimensions are observed [4,5].

The difficulty that traditional clustering algorithms encounter in dealing with this challenge has inspired the invention of subspace clustering, or projected clustering, which has been studied extensively in recent years. The goal of subspace clustering is to locate clusters with their own associated dimensions that are embedded in different subspaces of the original data space [4,5]. Based on the ways that the cluster subspaces are determined, subspace clustering can be generally classified into two main categories: hard subspace clustering and soft subspace clustering [5,6]. In this paper, we focus on soft subspace clustering, which measures the importance of each dimension to a particular cluster in the clustering process by automatically assigning different weightings to different dimensions. Soft subspace clustering algorithms can be grouped into two categories: *f*uzzy *w*eighting *s*ubspace *c*lustering (FWSC) [7] and *e*ntropy *w*eighting *s*ubspace *c*lustering (EWSC) [6]. Both of them use batch processing to update cluster centers, only after all data points' membership degrees have been obtained. However, in real-life environments, streaming data continuously arrives as being collected with chunks at different time intervals. Moreover, it may not be possible to retain huge amounts of data in the main memory due to the memory limit [8,9]. To address these problems, it is necessary to generate an effective evolving version of soft subspace clustering which has the ability to incrementally update models for high dimensional data streams.

During the last few years, a number of research efforts to cluster streaming data have been made. One traditional way to replace the batch update of cluster centers is based on an online learning strategy, and a large number of competitive

* Corresponding author. Tel.: +86 21 3420 2023.
E-mail address: wxzhulin@yahoo.com.cn (L. Zhu).

1568-4946/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.asoc.2013.03.002

learning-based online clustering algorithms [10–12] have been introduced. Banerjee and Ghosh combined frequency sensitive competitive learning and introduced an effective clustering technique called frequency-sensitive spherical *K*-means (fs-spkmeans) [10]. The online spherical *K*-means (OSKM) algorithm [12] investigates an online version of spherical k-means algorithm to cluster large text datasets. In addition, the extension of learning vector quantization [11] was developed based on the Winner-Take-More scheme that updates the recursion formulas of a cluster center by incorporating them with the membership degree of each cluster center. In general, the online version of clustering algorithms can not only achieve significantly better results than batch algorithms but can also incrementally update the cluster models with different learning strategies, which is an important and effective way of analyzing data streams, building on the understanding of the underlying distribution of data flow. By considering the advantages of competitive learning theory, we first introduce two online soft subspace clustering algorithms, OFWSC and OEWSC, which adopt an online learning strategy to modify the traditional batch soft subspace clustering approaches.

Although performance studies have shown that using an online clustering technique can achieve significantly better results than the batch version, it is still necessary to go through all data points several times, making it impossible to store large data streams. Recently, scalable clustering has been proposed to partition large-scale data or streaming data, by means of dividing data points into chunks and processing them continuously, segment by segment [9,13–17]. Bradley et al. first presented a scalable clustering technique [13], called ScaleKM, which selectively retains or compresses the samples that are important and discards the insignificant samples. By using a simple compression scheme, Farnstrom et al. proposed a simplification of ScaleKM [14], the simple single pass *K*-means algorithm, for large databases. In addition to these works addressing the crisp case of scalable clustering, Hall et al. generalized the scalable strategy to the fuzzy case and introduced two fuzzy scalable clustering algorithms: single-pass fuzzy *C*-means (SPFCM) [15] and online fuzzy *C*-means (OFCM) [16,17]. In summary, scalable clustering can not only handle streaming data efficiently, but can also produce partitions as good as online and batch subspace clustering methods. Motivated by these ideas, we propose evolving soft subspace clustering, which identifies clusters by assigning various weighting values to different dimensions of clusters in an incremental, step-wise clustering procedure. Specifically, by extending the online soft subspace clustering, we present two streaming soft subspace clustering algorithms, SSSC_F and SSSC_E, for partitioning data streams and revealing the important local subspace characteristics in data stream clustering procedures.

To the best of our knowledge, the proposed evolving soft subspace clustering is the first work to address the synthesis of the merits of soft subspace clustering with the beneficial properties of an online learning scheme, as well as providing a scalable subspace clustering strategy for streaming data. Comprehensive performance studies demonstrate that OFWSC and OEWSC are more effective in subspace clustering than the traditional batch methods, and SSSC_F and SSSC_E can obtain data stream clustering results as good as online and batch algorithms. The rest of this paper is organized as follows. In Section 2, a brief overview of existing approaches for soft subspace clustering is given. Section 3 provides a detailed description of online soft subspace clustering, OFWSC and OEWSC. The proposed streaming soft subspace clustering algorithms, SSSC_F and SSSC_E, are presented in Section 4. In Section 5, the clustering performance of four evolving subspace clustering methods on both synthetic and real-life datasets in comparison with other clustering techniques is reported. Finally, conclusions and future work are discussed in Section 6.

2. Soft subspace clustering

Soft clustering aims to group a set of given data points $X = \{x_1, x_2, \ldots, x_N\} \subset R^D$ into a set of clustering centers $V = \{v_i, 1 \le i \le C\}$. Let u_{ij} denote the membership degree of x_i belonging to v_i, then we can define the fuzzy *C*-partition matrix *U* of the given dataset, i.e., $U = \{u_{ij} | 1 \le i \le C, 1 \le j \le N\}$. To discover clusters from different subspaces, it is vital that a soft clustering algorithm has the capability to cluster data points by automatically weighting features in its clustering procedure. For this reason, a weighting w_{ik} is assigned to each dimension based on the importance of the *k*th dimension to the formation of the *i*th cluster. The subspaces of clusters can be identified by the weighting matrix $W = \{w_{ik} | 1 \le i \le C, 1 \le k \le D\}$ after soft subspace clustering [2,6,7].

2.1. Fuzzy weighting subspace clustering

*F*uzzy *w*eighting *s*ubspace *c*lustering (FWSC) [7] seeks to find clusters from fuzzy weighting subspace. In all the fuzzy weighting subspace clustering algorithms, a fuzzy weighting w_{ik}^{τ} is assigned to each feature of clusters with a fuzzy weighting index τ. The objective function of FWSC is generally formulated as:

$$J_{\mathrm{FWSC}} = \sum_{j=1}^{N}\sum_{i=1}^{C} u_{ij}^{m} \sum_{k=1}^{D} w_{ik}^{\tau}(x_{jk} - v_{ik})^2 \tag{1}$$

$$\text{s.t.} \quad \begin{array}{ll} 0 \le u_{ij} \le 1, & \sum_{i=1}^{C} u_{ij} = 1 \\ 0 \le w_{ik} \le 1, & \sum_{k=1}^{D} w_{ik} = 1 \end{array}.$$

By minimizing (1) using Lagrange multipliers, the updating equations for estimating center v_i, fuzzy weighting w_{ik} and membership degree u_{ij} can be derived by the theorem below.

Theorem 1. Assume $m > 1$ and $\tau > 1$, the necessary condition for the minimum of the objective function of FWSC in (1) yields the following update equations:

$$u_{ij} = \frac{(d_{ij})^{-1/m-1}}{\sum_{s=1}^{C}(d_{sj})^{-1/m-1}} \tag{2}$$

$$d_{ij} = \sum_{k=1}^{D} w_{ik}^{\tau}(x_{jk} - v_{ik})^2$$

$$v_{ik} = \frac{\sum_{j=1}^{N} u_{ij}^{m} x_{jk}}{\sum_{j=1}^{N} u_{ij}^{m}} \tag{3}$$

$$w_{ik} = \frac{(q_{ik})^{-1/\tau-1}}{\sum_{s=1}^{D}(q_{is})^{-1/\tau-1}} \tag{4}$$

$$q_{ik} = \sum_{j=1}^{N} u_{ij}^{m}(x_{jk} - v_{ik})^2$$

2.2. Entropy weighting subspace clustering

The entropy concept, which is used to represent the certainty of dimensions in the identification of a cluster, is also introduced into soft subspace clustering. Since the weightings in the extended subspace clustering methods are controllable by entropy, this type of algorithm is referred to as *e*ntropy *w*eighting *s*ubspace *c*lustering