Information Sciences 246 (2013) 133–154

Contents lists available at SciVerse ScienceDirect

Information Sciences

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

Fuzzy partition based soft subspace clustering and its applications in high dimensional data

Jun Wang [a], Shitong Wang [a,b,*], Fulai Chung [a], Zhaohong Deng [a]

[a] School of Digital Media, Jiangnan University, Wuxi, China
[b] Department of Computing, Hong Kong Polytechnic University, Hong Kong

ARTICLE INFO

Article history:
Received 18 January 2012
Received in revised form 14 May 2013
Accepted 20 May 2013
Available online 28 May 2013

Keywords:
Fuzzy clustering
Soft subspace clustering
Convergence
High dimensional data

ABSTRACT

As one of the most popular clustering techniques for high dimensional data, soft subspace clustering (SSC) algorithms have been receiving a great deal of attention in recent years. Unfortunately, most existing works do not cluster high dimensional sparse data and noisy data in an effective manner. In this study, a novel soft subspace clustering algorithm called PI-SSC is proposed. By introducing a partition index (PI) into the objective function, a novel soft subspace clustering algorithm that combines the concepts of hard and fuzzy clustering is proposed. Furthermore, the robust property of PI-SSC is analyzed from the viewpoint of ε-insensitive distance. A convergence theorem for PI-SSC is also established by applying Zangwill's convergence theorem. The results of the experiment demonstrate the effectiveness of the proposed algorithm in high dimensional sparse text data and noisy texture data.

1. Introduction

Clustering is an unsupervised learning process, the aim of which is to partition unlabeled data into different clusters so that similar data are assigned to the same cluster while dissimilar data are assigned to different clusters. Various clustering methods have been developed, including fuzzy clustering, spectral clustering, and hierarchical clustering, and they have been widely applied in many areas such as data mining, image segmentation, and time series analysis [12,17,22,29–32,42–46].

Most of the conventional clustering techniques utilize the Euclidean distance measure, which considers each feature of the dataset equally. This can be good enough if all or most features are equally important to every cluster. In many cases, however, features are not considered to be of equal importance and, accordingly, feature weights are often incorporated into computations of distance. In this way, the performance of clustering algorithms can be improved accordingly [33]. Recently, the development of feature weighting algorithms has become a hot topic in cluster analysis, with the basic idea being to assign each feature a weight for the entire dataset when calculating the distance between two data points. Among the feature weighting clustering algorithms, WFCM [31,32] and W-*k*-means [11] are two representatives. In WFCM feature weighting is performed prior to the clustering procedure, while in W-*k*-means it is integrated into the clustering procedure. Generally speaking, feature weighting algorithms learn a set of weights for the features involved and assign large values to important ones. Thus, they always obtain better results than the conventional clustering algorithms, which employ the traditional Euclidean distance measure.

The limitation of the feature weighting approach lies in the fact that the corresponding clustering algorithms are still based on all features and all clusters within a dataset sharing an identical feature space. For high dimensional datasets,

* Corresponding author at: School of Digital Media, Jiangnan University, Wuxi, China. Tel.: +86 13182791468.
E-mail address: Shwangst@yahoo.com.cn (S. Wang).

0020-0255/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ins.2013.05.029

however, different clusters are often correlated with different subsets of features, i.e., clusters may exist in different subspaces that consist of different subsets of features [4,10]. Subspace clustering, which has been extensively studied in recent years, is one effective data mining tool for datasets of this kind. The goal of subspace clustering is to locate clusters in different subspaces of a dataset. According to the ways in which the subspaces are identified, subspace clustering can be classified into two categories. One is referred to as hard subspace clustering, which is the identification of the exact subspaces for different clusters [1–3,23,24]. In this category of subspace clustering algorithms, the membership of a feature belonging to one cluster is identified by a binary value. The other category of subspace clustering is referred to as soft subspace clustering, in which data objects are grouped in the entire data space but different weighting values are assigned to different dimensions of clusters based on the importance of the dimensions in identifying the corresponding clusters [5,6,13,14]. Soft subspace clustering can be considered to be an extension of feature weighting clustering [11,31]. However, there is an obvious difference between them in that soft subspace clustering techniques assign a different vector of feature weights to each cluster, while feature weighting clustering techniques often take one common vector of feature weights over the whole dataset. This improvement makes soft subspace clustering more suitable for datasets with different clusters correlated with different subsets of features.

Although many soft subspace clustering algorithms have been successfully applied in various areas, their performance can be further improved. Currently, most of these algorithms have been developed based on the hard partition of a dataset [5,13,14] and can successfully deal with high dimensional data. However, most of them still fall short with respect to high dimensional sparse data and/or noisy data, such as text data and texture image data with noise. In this paper, we propose a novel soft subspace clustering algorithm called PI-SSC. Incorporating the partition index (PI) and a feature weighting metric with fuzzy weights into the objective function, the concept of cluster cores is introduced and the ideas of fuzzy and hard clustering are combined. The introduction of the partition index makes the proposed algorithm suitable for high dimensional sparse data and insensitive to noise in datasets.

The rest of the paper is organized as follows. In Section 2, some representative soft subspace clustering techniques are reviewed. In Section 3, we propose our PI-SSC algorithm and study its properties. In Section 4, we prove its convergence by using Zangwill's convergence theorem A. In Section 5, the experimental results are reported, indicating the advantage of PI-SSC over some representative clustering algorithms.

2. Related works

Recently, many soft subspace clustering algorithms have been proposed. Generally speaking, most of them can be said to deal with the problem of sfinding the local minimum of the following objective function

$$J(\mathbf{U},\mathbf{W},\mathbf{V}) = \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}^{m}\sum_{h=1}^{s} w_{ih}^{\alpha}(x_{kh}-v_{ih})^2 + \mathrm{H} \tag{1a}$$

under the constraints

$$\sum_{i=1}^{c} u_{ik} = 1, \quad u_{ik}\in[0,1] \tag{1b}$$

$$\sum_{h=1}^{s} w_{ih} = 1, \quad w_{ih}\in[0,1] \tag{1c}$$

where c is the number of clusters, n is the number of data points and s is the number of features. In Eq. (1a), the first term $\sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}^{m}\sum_{h=1}^{s} w_{ih}^{\alpha}(x_{kh}-v_{ih})^2$ can be interpreted as the total sum of all of the weighted distances of each data point to cluster centers, while the second term H is a penalty term that helps to generate a meaningful partition and effective soft subspaces. The parameter m is important and it indeed influences the fuzziness of the partition. When it takes the value of one, the algorithm is a k-means type of algorithm that is built on the hard partition of the dataset [5,7,13,14]. On the other hand, when $m > 1$, the algorithm is a fuzzy clustering algorithm that is built on the fuzzy partition of the dataset [6]. The parameter α is another important parameter that is closely related to how the subspace is weighted. According to different approaches to subspace weighting, soft subspace clustering can be divided into two categories: entropy weighting subspace clustering and fuzzy weighting subspace clustering. In entropy weighting subspace clustering, the parameter α always takes the value of one and in fuzzy weighting subspace clustering, the constraint $\alpha > 1$ should be satisfied.

2.1. Fuzzy weighting subspace clustering

The AWA algorithm proposed by Chan et al. [5] may be the first fuzzy weighting subspace clustering algorithm. Its objective function is formulated as follows: