Information Sciences 381 (2017) 104–123

Contents lists available at ScienceDirect

Information Sciences

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

A clustering algorithm for stream data with LDA-based unsupervised localized dimension reduction

Sirisup Laohakiat*, Suphakant Phimoltares, Chidchanok Lursinsap

Department of Mathematics and Computer science, Chulalongkorn Univeristy, Thailand

ARTICLE INFO

Article history:
Received 6 January 2016
Revised 14 October 2016
Accepted 25 November 2016
Available online 27 November 2016

Keywords:
Stream data clustering
Dimension reduction
Linear discriminant analysis
Clustering algorithm
Linear discriminant analysis subspace

ABSTRACT

We present an algorithm for clustering high dimensional streaming data. The algorithm incorporates dimension reduction into the stream clustering framework. When a new datum arrives, the algorithm performs dimension reduction to find a local projected subspace using unsupervised LDA (Linear Discriminant Analysis)-based method. The obtained local subspace would maximally separate the nearby micro-clusters with respect to the incoming point. Then, the incoming point is assigned to a micro-cluster in the projected space, rather than in the full dimensional space. The experimental results show that the proposed algorithm outperforms its counterpart streaming clustering algorithms. Moreover, when compared with traditional clustering algorithms which require the whole data set, the proposed algorithms shows comparable clustering performances with much less computation time for large data sets.

1. Introduction

Recently, high dimensional data have been generated and collected in many applications. Various analyzing and mining methods for high dimensional data have been proposed [12,13,18,28]. Among these studies, clustering high dimensional data has been one of the highly addressed problems due to its important role in data mining. The challenge of clustering high dimensional data is the curse of dimensionality, one aspect of which is that the differences of distances between pairs of data points become less distinct [6], leading to difficulties in detecting clusters in the data set. Traditionally, a linear dimensionality reduction technique, such as PCA (Principal Component Analysis) [24], LDA (Linear Discriminant Analysis) [17], or any feature selection algorithms such as [44], is performed to reduce the dimensions of the data first, then the data with less dimensions can be clustered with ordinary clustering algorithm. In this approach, the dimension reduction is conducted globally on the entire data set, leading to a single set of the reduced dimensions. However, when the data contain nonlinear structure, using one set of the reduced dimensions to represent the original data could be insufficient.

Recently, there have been several studies using local linear models to represent data with nonlinear structure. For example, LLE uses local linear embeddings with low dimensionality to represent high dimensional data with nonlinear structure [39]. Li et al. proposed a localized feature selection algorithm which uses different subsets of features for data in different clusters [27]. Using several reduced-dimension linear models to represent the original feature space can improve the efficiency and flexibility of the dimensionality reduction.

* Corresponding author.
E-mail address: sirisup.l@gmail.com (S. Laohakiat).

http://dx.doi.org/10.1016/j.ins.2016.11.018
0020-0255/

Based on the same rationale of localized dimensional reduction, projected clustering algorithms, for example PreDeCon and PROCLUS, incorporate both dimension reduction and clustering together [25] by grouping data points into clusters which are defined in projected subspaces, rather than the original full dimensional space. Different clusters could lie in different subspaces defined by distinct subsets of features. In this sense, these algorithms adopt the idea of localized dimension reduction. However, Ding et al. [14] have argued that, using a subset of original features to represent the projected subspace restricts the resulting subspace to be axis-parallel, whereas the actual subspaces are not necessarily limited to those parallel to the axes. This limitation could impair the flexibility of the dimensionality reduction. As a result, they have proposed the use of LDA subspace, in which the clusters in the subspace are well separated while the resulting subspace is not limited to those parallel to the axes. Integrating LDA and clustering algorithm into a single framework has been investigated in many studies, for example LDA-Km [14], DCA [46], and Discriminative K-means [49].

Another well-known technique that can handle high dimensional data sets efficiently is locality sensitive hashing (LSH) [41]. LSH has been proposed by Indyk and Motwani [23] for finding approximate nearest neighbors in high dimensional data set. Based on the idea that LSH would create hash functions which map closed data points to the same bucket, LSH can be found in many clustering applications, for example web pages clustering [20]. However, in this study, we focus on the technique that performs feature reduction, rather than those that use all features of the data set like LSH.

All of the aforementioned algorithms have been designed to handle ordinary data set in which the entire data set is available before the clustering process. Recently stream data, continuously generated data, have collected and analyzed from many real-time monitoring systems. Stream data differ from the traditional data sets in several aspects. For example, data points of stream data would accumulate over time, rather than being available at the beginning like the traditional data sets. Moreover, the volume of stream data is usually large due to the continuous accumulation of data. Several algorithms for clustering stream data have been proposed in recent literature, for instance, single pass K-median clustering [10,19,36], single pass density-based clustering [9,11,26], streaming K-means [8], evolving vector quantization [29]. In recent works, concept drift is also taken into consideration such as [4,42]. Application of clustering stream data can be found in several works, for example [32].

With the high throughput data acquisition systems, data streams with high dimensionality have been more widespread, leading to the requirement of clustering algorithms for high dimensional stream data. Several algorithms have been proposed for example HPStream [3] and HDDStream [35]. HDDStream adopts the idea of PreDeCon [7] in assigning weights to each feature in calculating distance between a data point and the existing clusters. HPStream performs feature selection for each cluster based on the value of the variance of each feature. These two algorithms perform dimension reduction based on feature selection which lead to the clusters being defined on axis-parallel subspaces. Despite being pointed out in [14] that LDA subspace is more suitable to perform clustering than axis-parallel subspace, to our knowledge, no algorithm for stream data which uses LDA subspace has been proposed.

Moreover, the existing algorithms perform global dimension reduction, resulting in one set of LDA subspace representing the entire data set. Adopting localized LDA should add more flexibility to the framework, and thus improves the clustering performance. As a result, we propose a clustering algorithm for stream data which integrates localized LDA and clustering into a single framework.

To address the two issues regarding the benefits of LDA subspace and the local representation of subspace, we design a stream data clustering algorithm with the following features.

1. We integrate dimension reduction projected on LDA subspace into a density based clustering algorithm for streaming data.
2. Instead of determining LDA subspace globally, the proposed algorithm uses localized LDA subspaces to improve the flexibility of dimension reduction.

The proposed algorithm, named as LLDstream, adopts density-based one-pass clustering method based on DenSteam [9] and DBSCAN [16]. LLDstream assigns data points to clusters in localized LDA subspace determined by an algorithm called ULLDA (unsupervised localized linear discriminant analysis). ULLDA uses the same rationale as LDA in local scale. Instead of finding a single subspace which globally maximizes class-separation as in traditional LDA, ULLDA determines the local LDA subspace for each incoming data point based on the local clusters near that point.

The clustering performance of LLDstream is evaluated in comparison with the state-of-the-art stream data clustering algorithms. Two of which are projected clustering algorithms designed for handling high dimensionality stream data. Performance evaluation also includes the comparison of LLDstream with the clustering algorithms for conventional data sets which require the entire data during clustering. Using one-pass clustering scheme, LLDstream can handle large data set more efficiently with satisfactory clustering results.

The rest of this paper is organized as follow. Section 2 gives an overview of the proposed algorithm. Section 3 summarizes the concept of LDA and presents the idea of unsupervised localized linear discriminant analysis. The detail of LLDstream is provided in Section 4. The comparative performances of our algorithm and other state-of-the-art algorithms are discussed in Section 5. Section 6 concludes this paper.