Neurocomputing 180 (2016) 16–26

Contents lists available at ScienceDirect

Neurocomputing

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

PTS: Projected Topological Stream clustering algorithm

Cássio M.M. Pereira [*], Rodrigo F. de Mello

Institute of Mathematical and Computer Sciences, São Carlos, SP 13566-590, Brazil

ARTICLE INFO

Article history:
Received 28 March 2015
Received in revised form
15 July 2015
Accepted 18 August 2015
Available online 10 November 2015

Keywords:
Data streams
Data clustering
Persistent homology
Topological data analysis

ABSTRACT

High-dimensional data streams clustering is an attractive research topic, as there are several applications that generate a high number of attributes, bringing new challenges in terms of partitioning due to the curse of dimensionality. In addition, those applications produce unbounded sequences of data which cannot be stored for later analysis. Although the importance of this scenario, there are still very few algorithms available in the literature to meet this task. Despite the theoretical foundation of mathematical topology for dealing with high-dimensional spaces, none of those approaches have investigated the problem of finding topologically similar projected clusters in high-dimensional data streams. Among the advantages of topology is the possibility to analyze data in a coordinate-free and noise-robust manner. In a previous research, we have shown that topologically similar clusters can be meaningful considering real-world data sets. In this paper, we extend those ideas and propose PTS, an algorithm for finding topologically similar clusters in high-dimensional data streams. The algorithm is capable of finding traditional projected clusters and then merging them according to topological features computed using persistent homology. Experiments with synthetic data streams of dimensions $d=8, 16, 32, 64$ and 128 confirm the ability of PTS to find topologically similar projected clusters.

1. Introduction

Subspace clustering of data streams [9] is an emerging research field, with few algorithms available for the task when compared to the traditional batch scenario. Despite of that, there are many applications which produce continuous high-dimensional data, such as in the medical and scientific fields. For instance, the LHC (Large Hadron Collider) at CERN alone generates 500 EB of data, i.e., 500 billion gigabytes, per day from 150 million servers.

Data streams algorithms are more useful in such scenarios, as they make the more realistic assumption that data evolves over time, i.e., the distributions governing data generation may change. Besides that, many important tasks consist in high-dimensional data, in which traditional distance measures become irrelevant, as the sparsity of data tends to make all points equidistant. Such limitations led to the development of new algorithms that can cope with these evolving and high-dimensional scenarios.

In such settings, many attributes are expected to be irrelevant to the analysis. Such attributes can hide the true clusters present in a subspace of the data stream. Besides irrelevant attributes, clusters may also be defined in different subsets of attributes, thus making traditional feature selection algorithms infeasible [12], as attribute relevance is local to clusters and not a global feature of the stream.

According to Moise et al. [16], subspace clustering techniques search for all clusters of points in all subspaces of a data set according to their respective cluster definition. On the other hand, projected clustering techniques define a projected cluster as a pair, which consists of a subset of data points and a subset of data attributes, such that the points are close when projected onto that particular subset of attributes and farther apart when projected onto the remaining attributes. In this paper, we are concerned with finding orthogonal projections that are statistically relevant to find the initial clusters, thus our proposed technique follows the projected clustering model. Formally, the task of finding projected clusters in high-dimensional data streams can be defined as follows. Let $\mathcal{D}=\{p_1, p_2, \ldots, p_\infty\}$ be a possibly-infinite data stream composed of d-dimensional points $p_i=\{p_{i1}, p_{i2}, \ldots, p_{id}\}$. A partition $\Gamma=\{\gamma_1, \ldots, \gamma_k\}$ consists of k clusters, such that $\gamma_k=(\Pi_k, A_k)$, in which Π_k is a subset of points and A_k is a subset of attributes, such that points in Π_k are closer together along attributes in A_k than to other points in the same or other dimensions. Although there are a few algorithms available for high-dimensional data stream clustering [1,17,4,9], none have explored the idea of discovering

[*] Corresponding author. Work address: Instituto de Ciências Matemáticas e de Computação - USP: Avenida Trabalhador São-carlense, 400 - Centro CEP: 13566-590 - São Carlos - SP - Brazil.
E-mail addresses: cpereira@icmc.usp.br (C.M.M. Pereira), mello@icmc.usp.br (R.F. de Mello).

http://dx.doi.org/10.1016/j.neucom.2015.08.094
0925-2312/© 2015 Elsevier B.V. All rights reserved.

topologically similar projected clusters, which confirms an important gap as mathematical topology provides a theoretical foundation for dealing with high-dimensional spaces. Topological data analysis (TDA) [3] aims to study not only quantitative, but especially qualitative aspects of data. For instance, while circles and squares are different geometric objects, they are the same mathematical object when it comes to topology, as one can be continuously deformed into the other. Topology has the advantage of studying properties of geometric objects that do not depend on the coordinate system, but rather on intrinsic geometric features of how objects are distributed in space, thus it is coordinate-free. One of the computational topology fields that presents interesting qualitative properties is persistent homology [24]. One of them, that can be estimated, are Betti numbers [24], which are used to discriminate topological spaces based on the connectivity of n-dimensional simplicial complexes. For instance, a p-simplex σ is the convex hull of $p+1$ linearly independent points $x_0, x_1, \ldots, x_p \in \mathbb{R}^d$. More intuitively, a 0-simplex is a vertex, a 1-simplex is an edge, a 2-simplex is a triangle, a 3-simplex is a tetrahedron, and so forth. Betti numbers indicate the number of d-dimensional holes present in a multi-scale analysis of the connectivity of simplicial complexes constructed from data. A 0-d hole counts the number of connected components, while a 1-d hole is the space inside a circle, a 2-d hole the space inside a balloon and so forth. Betti numbers can be used as shape descriptors that are stable under shape deformations, thus they are useful, for example, when working on noisy shapes in content-based image retrieval [10]. We have previously shown that topologically similar clusters can be meaningful to partition objects [18]. In this paper, we extend those ideas and propose a clustering algorithm for the task of finding topologically similar projected clusters in data streams. The algorithm is named PTS and performs a pipeline that consists in finding relevant projected areas of space, which are then merged together according to their topological properties, estimated via Betti numbers. Fig. 1 shows the pipeline performed by our approach. We present experiments on synthetic data streams that confirm the ability of our technique to retrieve topologically similar projected clusters in streaming scenarios.

2. Background concepts and related work

2.1. Background concepts on computational topology

Carlsson [3] points out that the clustering task is usually formulated with many ambiguities, such as how several parameters are arbitrarily determined, e.g., the density threshold ϵ for DBSCAN or the number of clusters K for K-Means. In light of this, one of the concerns of topological data analysis is to understand how geometric objects relate to one another when constructed with varying parameter values, instead of an arbitrary setting. This relates to the mathematical idea of functoriality, which indicates that invariants should also be related to the maps among objects and not just among the objects themselves. That concept is applied in our proposal through the use of persistent homology, for which we give a brief introduction next. Persistent homology embraces this idea by analyzing how point connectivity changes with varying distance values. For a detailed introduction to concepts on computational topology, we refer the reader to Zomorodian [26], Carlsson [3], and Holzinger [10].

Formally, a topology on a set X is a subset $T \subseteq 2^X$ such that (1) if $S_1, S_2 \in T$, then $S_1 \cap S_2 \in T$; (2) if $\{S_j | j \in J\} \subseteq T$, then $\cup_{j \in J} S_j \in T$; and (3) $\varnothing, X \in T$. This implies that topology is a system of subsets that describe the connectivity of a set.

The pair (X,T) of a set X and a topology T forms a *topological space*. In data mining, we are used to metric spaces, specially the Euclidean space. A metric space is a topological space endowed with a metric function d. The Euclidean space can thus be defined as the Cartesian product of n copies of $\mathbb{R}$ along with the Euclidean metric: $d(x,y) = \sqrt{\sum_{i=1}^{n} (u_i(x) - u_i(y))^2}$, where u_i is the i-th Cartesian coordinate function, thus forming the n-dimensional Euclidean space $\mathbb{R}^n$.

In persistent homology, we ultimately want to compare topological spaces based on the characteristic holes that they encompass. Because we usually operate with finite point clouds in a metric space, we first need to discretize the space in order to add the notion of connectivity, i.e., topology. That is done through the creation of simplicial complexes.

A p-simplex σ is the convex hull of $p+1$ linearly independent points $x_0, x_1, \ldots, x_p \in \mathbb{R}^d$ [24]. More intuitively, a 0-simplex is a vertex, a 1-simplex is an edge, a 2-simplex is a triangle, a 3-simplex is a tetrahedron, and so forth. A simplicial complex K is a finite set of simplices such that for $\sigma \in K$, all of its faces are also in K.

The core idea in persistent homology is to analyze how holes appear and disappear, as simplicial complexes are created. To do that, a filtration is constructed. An increasing sequence of ϵ values, i.e. distance values, produces a filtration, such that a simplex enters the sequence no earlier than all its faces. A common way to do this is by using Vietoris–Rips complexes [24], that are only added to the filtration at $\epsilon = \epsilon'$ if the distance between two points in σ is less than or equal to ϵ'. This is useful since all simplices that are present in ϵ' will also be contained in ϵ'', if $\epsilon'' \geq \epsilon'$.

To illustrate these concepts, consider the point cloud in a 2-d metric space illustrated in Fig. 2, consisting of a five point set:

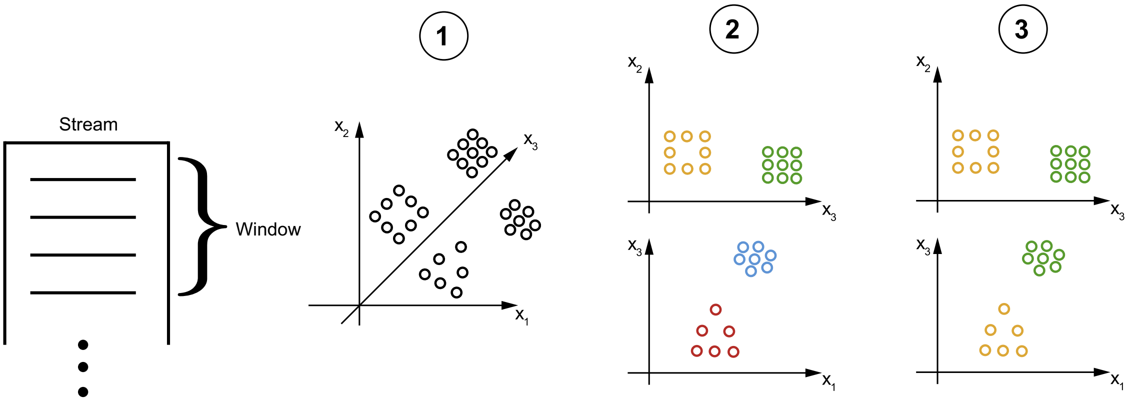

Fig. 1. Pipeline of our approach. In step 1, a window of the data stream is obtained for analysis. In step 2, traditional projected clusters are found, along with their relevant dimensions. In step 3, each projected cluster is discretized so a topological space can be inferred from connectivity information. The Betti numbers of each discretized space are then used to merge clusters based on their topological features. In the example, the empty square and triangle are merged since both surround a hole in the middle.