Computer Vision and Image Understanding 121 (2014) 21–26

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

Towards topological analysis of high-dimensional feature spaces

Hubert Wagner [a,*], Paweł Dłotko [a,b]

[a] Institute of Computer Science, Jagiellonian University, Lojasiewicza 6, Krakow, Poland
[b] Department of Mathematics, University of Pennsylvania, Philadelphia, PA 19104-6395, USA

ARTICLE INFO

Article history:
Received 30 September 2012
Accepted 18 January 2014

Keywords:
Computational topology
Persistent homology
Rips complex
Text mining
Feature space

ABSTRACT

In this paper we present ideas from computational topology, applicable in analysis of point cloud data. In particular, the point cloud can represent a feature space of a collection of objects such as images or text documents. Computing persistent homology reveals the global structure of similarities between the data. Furthermore, we argue that it is essential to incorporate higher-degree relationships between objects. Finally, we show that new computational topology algorithms expose much better practical performance compared to standard techniques.

© 2014 Elsevier Inc. All rights reserved.

1. Motivation

The purpose of this paper is to introduce concepts and techniques from computational topology in the context of image understanding and pattern recognition. We think that using such methods in conjunction with the standard tools present in these fields, can give rise to new effective solutions.

Faced with a large collection of documents, including images, it is useful to have a global view of this dataset. In this paper we argue that tools of computational topology can be used to capture topological structure of point-cloud data and that this information can be useful. In particular, using persistent homology we are able to capture global, higher-dimensional patterns within a feature space.

Data mining methods often use graph-theoretical approaches [15]. Analysing the connected components of the graph of *similarities* between pairs of objects is a simple example. From a topological perspective, such analysis operates on 1-dimensional *complexes* (only pairs of documents are considered) and gives 0-dimensional topological information.

Higher dimensional relationships, i.e. relationships between larger subsets of data, are sometimes used in data-mining. For example, the number of triangles (3-cliques) is an important descriptor of the connectivity of a social or collaborative network [10]. Rather than finding just the *number* of such higher-dimensional elements, we would like to compute their topological structure.

We believe that mining a higher dimensional *topological structure* within a set of objects can give an important insight into the data. For example, [3] shows that data coming from natural images form a topological Klein bottle.

The original motivation for our project was an application from the area of text-mining, as described in [21]. The following paper serves as an extension, putting these techniques in a wider context. To be specific, we show the applicability of this framework in the context of computer vision and image understanding. Additionally, we update some of the information contained in the previous paper [21], based on recently gained experience.

We intend to accomplish three goals. First, introducing a higher-dimensional similarity measure, we show that the point-cloud can be interpreted as a simplicial complex, with meaningful filtration values defined on simplices of all dimensions. Such a representation allows for treatment with topological tools. Second, we argue that a higher-dimensional analysis, based on topological tools, can indeed be interesting and relevant. Third, we demonstrate that using recent algorithmic techniques, much larger datasets can now be handled.

2. Input data

By a *feature space* we mean a vector space, where each coordinate corresponds to the numerical value of a certain feature. Each possible tuple of features' values can be represented as a point (vector) in this space. Depending on the application, these points may correspond to images, text documents, etc.

In general, our input consists of a set of objects (images, text documents, etc.). We also choose a certain *similarity measure*,

* Corresponding author.
E-mail addresses: wagner@ii.uj.edu.pl (H. Wagner), dlotko@sas.upenn.edu (P. Dłotko).

http://dx.doi.org/10.1016/j.cviu.2014.01.005
1077-3142/© 2014 Elsevier Inc. All rights reserved.

quantifying how related, or similar, objects are. Normally, similarity is a pairwise function. In our methodology, we consider higher-dimensional relationships, that is the relationships within sets of arbitrary size. In contrast, standard, graph-theoretical methods capture only pair-wise relationships, effectively operating on a *one-dimensional* structure.

The values of similarity range from 0 (completely unrelated objects) to 1 (indistinguishable objects). Therefore, in our method we identify objects with similarity equal to 1.

Importantly, we also define dissimilarity, $dsim(\cdot) = 1 - sim(\cdot)$, where *sim* is a similarity measure. Dissimilarity fits conveniently with the framework of persistent homology. Note that dissimilarity it is not necessarily a metric, but we do require $dsim(x,x) \leqslant dsim(x,y) = dsim(y,x)$.

3. Computational topology

In this section we give a brief introduction to computational topology. For a formal introduction see [8]. A paper by Carlson [3] is an important work, which shows that analysis of higher-dimensional data can be meaningful. A number of papers dealing with topological analysis in lower-dimensional spaces exist, but these techniques are hard or impossible to generalize to higher dimensions [16]. A recent paper by Zomorodian [20] deals with building Rips complexes of high dimensional data.

A finite collection of finite sets, S, is an abstract *simplicial complex* if for every $t \in S$ and for every $s \subset t$ we have $s \in S$. Every element $t \in S$ is a *simplex* and its *dimension* is defined as $card(t) - 1$. By S_k we denote the k-skeleton of complex S, i.e. all simplices in S with dimension $\leqslant k$. If $s \subset t$ and $card(t) - card(s) = 1$, we say that s is a *face* of t and t is a co-face of s. *(Co-)boundary* is the set of all (co-)faces of a simplex. A simplex of dimension 0, 1, 2, 3 is respectively: *a vertex, an edge, a triangle* and *a tetrahedron*.

We outline the computations performed: Starting from the point-cloud equipped with a *dissimilarity measure* we construct a *filtered simplicial complex*, which encodes higher dimensional topological information, and can be viewed as a higher-dimensional analog of a graph. Then, we compute the *persistence diagram* which encapsulates *persistent homology* on this data. We now proceed to define and explain these concepts.

3.1. Čech and Rips complexes

A point cloud, can be imagined as a sample of an underlying space. We reconstruct this space as a union of balls of a certain radius. (Later we will use persistent homology, so instead of fixing this radius, it will become a parameter.) There are two standard constructions, which yield a combinatorial representation of such a union of balls, in the form of a simplicial complex. Let $B_x(r)$ denote a ball centered at x with radius r. For a given point cloud P, we define the Čech complex:

$$\check{\text{C}}\text{ech}(r) := \left\{ \sigma \subseteq P | \bigcap_{x \in \sigma} B_x(r) \neq \emptyset \right\}.$$

Similarly, we define the Rips complex:

$$\text{Rips}(r) := \{ \sigma \subseteq P | max_{a,b \in \sigma} d(a,b) < 2r \}.$$

For sufficiently nice spaces, such as $\mathbb{R}^n$ with the standard topology, Čech complex has the homotopy type of the union of balls. In particular, it has the same homology [8]. Rips complex, being easier to compute, is usually used in practice, even though it might include some spurious topological information. In our setting the space might be more exotic, depending on the chosen dissimilarity measure, and in general the Čech property might not hold. Still, this is a useful intuition.

3.2. Persistent homology

Homology is a mathematical formalism used to define and identify basic topological features, called *holes*. Holes are defined for arbitrary dimensions, and in three ambient dimensions they are intuitive: 0-dimensional holes are related to the gaps between connected components, 1-dimensional ones can be viewed as tunnels (like a hole in a donut). 2-dimensional holes are cavities (inside of a balloon). See [8] for a formal definition of homology. By *homology class* we simply mean an individual hole.

Persistent homology describes the changes in homology when a certain scale parameter is varied. This way it can be viewed as a multi-scale view of topology. More formally, given a simplicial complex $\mathcal{K}$ and a filtering function $g : \mathcal{K} \to [0,1]$, *persistent* homology studies homological changes of the sub-level complexes: $\mathcal{K}_t = g^{-1}([0,t])$. Changing t from 0 to 1 induces a sequence of complexes called a filtration. The complex with the filtering function is called a filtered complex. Importantly, we require that $g(a) \leqslant g(B)$, whenever a is a face of B, which we call the *filtration property*. It implies that for every $t \in \mathbb{R}$, $\mathcal{K}_t$ is a *complex*, namely a simplex appears no sooner than its faces in a filtration.

Persistent homology captures the birth and death times of homology classes of the sub-level complexes, as t changes from 0 to 1. By birth, we mean that a homology class is created; by death, we mean it either becomes trivial or becomes identical to some other class born earlier. The *persistence*, or lifetime of a class, is the difference between the death and birth times.

Fig. 1 shows three levels of a filtration, built over a point cloud simpled from a figure eight. At consecutive filtration levels, holes are created and then killed. We start from a number of connected components. In the middle level, the connected components merge, and a small 1-dimensional cycle is formed. This cycle (outlined by the red ellipse), however has small *persistence*, because it is immediately glued-in as the balls grow. Then two 1-dimensional cycles are created. If we changed the parameter further, the cycles would also be glued-in, but they persisted over a large change of parameter. We can identify three homology classes with relatively high persistence, namely the unique connected component, which persists forever, and the two one-dimensional cycles. Clearly, these homological features correspond to the apparent shape of figure eight.

An important justification for the usage of persistence is the stability theorem. Cohen-Steiner et al. [7] proved that putting some mild assumptions on two filtering functions f and g from $\mathcal{K}$ to $[0,1]$, the so-called bottleneck distance (d_B, see [7]) between the persistence of $\mathcal{K}$ filtered by function f (denoted as $H^p(\mathcal{K},f)$) and

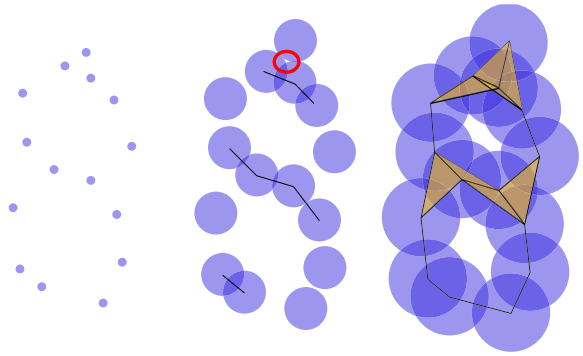

Fig. 1. Example of a point cloud, and a union of balls of growing size. A simplicial complex, called Rips complex is overlaid, being a combinatorial representation of the union of balls. Persistent homology captures the changes in homology for the growing radius of balls.

Download English Version:

https://daneshyari.com/en/article/525677

Download Persian Version:

https://daneshyari.com/article/525677

Daneshyari.com