Pattern Recognition 48 (2015) 391–401

Contents lists available at ScienceDirect

Pattern Recognition

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

An entropy-based persistence barcode ☆

Harish Chintakunta [a], Thanos Gentimis [a], Rocio Gonzalez-Diaz [b,*], Maria-Jose Jimenez [b], Hamid Krim [a]

[a] ECE Dept., NCSU, Raleigh, NC USA
[b] Applied Math Dept., School of Computer Engineering, U. Seville, Spain

ARTICLE INFO

Article history:
Received 1 October 2013
Received in revised form
27 June 2014
Accepted 28 June 2014
Available online 8 July 2014

Keywords:
Persistent homology
Persistence barcodes
Hasse diagram
Simplicial complexes
Entropy
Filtration
Filter

ABSTRACT

In persistent homology, the persistence barcode encodes pairs of simplices meaning birth and death of homology classes. Persistence barcodes depend on the ordering of the simplices (called a filter) of the given simplicial complex. In this paper, we define the notion of "minimal" barcodes in terms of entropy. Starting from a given filtration of a simplicial complex K, an algorithm for computing a "proper" filter (a total ordering of the simplices preserving the partial ordering imposed by the filtration as well as achieving a persistence barcode with small entropy) is detailed, by way of computation, and subsequent modification, of maximum matchings on subgraphs of the Hasse diagram associated to K. Examples demonstrating the utility of computing such a proper ordering on the simplices are given.

1. Introduction

In recent years persistent homology has successfully been used to characterize topological properties of given sets and structures [10]. Persistence and ZigZag barcodes [3] have helped identify characteristic properties of the underlying space in data mining, network coverage and social networks. The theory of persistence has proven to be robust and invariant under small perturbations of the original state of the data. The algebraic structure of the set of persistence barcodes viewed as a module is well understood [27], and various stability theorems are proven in [4] when the space is equipped with the Wasserstein distance. In particular, it is proven that the persistence barcode is stable under "noise" that is, small changes in the function used to create a filtration to compute persistent homology imply only small changes in the persistence barcode.

It is obvious that a persistence barcode depends heavily on the filter considered for such computation. An "interval" in a persistence barcode representation is a horizontal line segment in a plane whose length is equal to the lifespan of the corresponding homology class. A typical persistence barcode contains short-lived intervals representing topological holes (homology classes) which may not represent real features of the space being analyzed. In such cases one has to set a "threshold" of significance on the length of the intervals of the persistence barcode, and this is normally carried out a posteriori (see, for example, [14] where authors consider a "simplified" barcode). On the other hand, if someone examines in depth typical cases where persistent homology is used, one will be faced with the inherent problem of "noise" in the persistence barcodes. A similar problem was posed in [6,12] where various algorithmic results were presented. Those problems take the form of non-significant topological holes in sensor networks and inefficiencies of the filtration coming from the construction of the Rips (or Čech) complex in data sets.

Motivated by practical applications of persistent homology computation, our starting point is a given simplicial complex and an initial filtration. Although the total number of intervals in a persistence barcode remains invariant (as we will see later), the lengths of its intervals depend on the selected filter. Since non-significant intervals (i.e., intervals with short length) may not imply relevant homological information, we are interested in looking for a filter which preserves the partial ordering imposed by the given filtration, in order to minimize the number of significant intervals and maximize their lengths.

From an information-theoretic viewpoint, and if we interpret the number of significant intervals as the coding length of a complex, our goal is to select the most "parsimonious" representation (also by Occam's razor principle). As is also well known, the coding length is intimately related to the notion of entropy (i.e., a

☆ Authors partially supported under Grant No. MTM2012-32706, DTRA Grant No. HDTRA1-08-1-0024, and NSF Grant No. CCF-1217874.

* Corresponding author.
E-mail addresses: hkchinta@ncsu.edu (H. Chintakunta), agentim@ncsu.edu (T. Gentimis), rogodi@us.es (R. Gonzalez-Diaz), majiro@us.es (M.-J. Jimenez), ahk@ncsu.edu (H. Krim).

http://dx.doi.org/10.1016/j.patcog.2014.06.023
0031-3203/

topological entropy of the barcodes in our case). While one would ideally want to balance the minimization with a penalty term of the number of non-significant intervals [18], further knowledge of statistical distribution of the long intervals is required and postponed to future work.

This paper completes the work proposed in [12] by providing additional insights, examples, results and proofs. A new definition of "minimal barcode" is given using the notion of entropy, and an algorithm for computing persistence barcodes with small entropy starting from a given filtration is provided.

The remainder of this paper is organized as follows. Section 2 covers the relevant background material. In Section 3, we give the definition of minimal barcodes based on the notion of entropy. In Section 4, an algorithm for computing a filter which preserves the partial ordering imposed by the filtration and whose associated persistence barcode has small entropy, is computed on the Hasse diagram of the poset of faces of a given complex. Section 5 is devoted to relations between minimal barcodes and discrete Morse theory. Conclusions and future work are presented in Section 6.

2. Preliminaries

Homology theory uses algebraic groups to encode the topological structure of a simplicial complex K. In general, we will always consider connected, finite simplicial complexes K (i.e., with m simplices, $m<\infty$) unless stated otherwise.

2.1. Homology

The set of i-simplices, σ^i (superscript denotes dimension), will be denoted by K^i. The number of simplices in a set S is denoted by $|S|$. The dimension of the simplicial complex is $n \geq 0$ if $K^n \neq \varnothing$ and $K^m = \varnothing, \forall m > n$. Finite formal sums of simplices of K^i with coefficients in a field (called *i-chains*), define an additive Abelian group structure on K^i. In our case, the coefficients for those sums belong to $\mathbb{Z}_2$ thus the group of i-chains, $C_i(K;\mathbb{Z}_2)$ is a vector space over $\mathbb{Z}_2$ with basis elements the simplices of dimension i.

If a simplex σ is a face of another simplex σ', we write $\sigma \prec \sigma'$. We say that σ' is a *coface* of σ. A *proper face* of $\sigma \in K^i$, is a face of $\sigma$of dimension $i-1$. The *boundary* of σ, denoted by $\partial(\sigma)$ is the formal sum (with coefficients in $\mathbb{Z}_2$) of the proper faces of σ. The boundary operator is extended to all chains of K by linearity.

An i-chain a is an *i-cycle* if $\partial_i(a)=0$ i.e., $a \in \operatorname{Ker} \partial_i$; it is an *i-boundary* if there is an $(i+1)$-chain b such that $\partial_{i+1}(b)=a$ i.e., $a \in \operatorname{Im}(\partial_{i+1})$. Two i-cycles a and a' are *homologous* if $a+a'$ is an i-boundary. Since $\partial_i\partial_{i+1}=0$, $\operatorname{Im}(\partial_{i+1}) \subseteq \operatorname{Ker} \partial_i$. The quotient of i-cycles over i-boundaries is the i-th *homology group of K* i.e., $H_i(K)=\operatorname{Ker}(\partial_i)/\operatorname{Im}(\partial_{i+1})$. The elements of $H_i(K)$ are called *homology classes*.

Since the considered field of coefficients is $\mathbb{Z}_2$, the *i-th Betti number* (denoted by β^i) is the rank of the i-th homology group of K.

Then, the basic topological structure of K is quantified by the number of independent classes in each homology group. See [25,15].

Given a simplicial complex K, a nested sequence of simplicial complexes

$$\varnothing = K_0 \subset K_1 \subset \cdots \subset K_{p-1} \subset K_p = K$$

is called a *filtration* of K. An ordering of the simplices of a simplicial complex $K=\{\sigma_1,\dots,\sigma_m\}$ is called a *filter* if it satisfies the property that $s<t$ whenever $\sigma_s \prec \sigma_t$. Then we can create a filtration by setting:

$$K_t = \{\sigma_1,\dots,\sigma_t\}, \quad \text{for } 1 \leq t \leq m.$$

2.2. Metric filtrations

Many applications of Computational Topology start with a cloud of points embedded in $\mathbb{R}^n$. Using a specific radius r one can then define Alpha-complexes $A(r)$, Čech complexes $C(r)$, and Rips-complexes $R(r)$. Furthermore, one can obtain a filtration

$$\varnothing = K_0 \subset K_1 \subset \cdots \subset K_{p-1} \subset K_p = K$$

by gradually increasing r where K_i is $A(r_i)$, $C(r_i)$ or $A(r_i)$ depending on the complex K we are creating (Alpha, Čech or Rips-complex) and $r_i < r_j$ if $i<j$ (see [7, page 70] and [10]). In particular, all vertices enter at K_0, and K_t and K_{t+1} differ by at least one simplex.

2.3. Lower-star filtrations

In lecture 11 of their course[1] in computational geometry and topology, Edelsbrunner and Kerber argue that if we choose a reasonable filtration, we can learn more about a complex than just analyzing its Betti numbers. They propose the creation of a filtration given some function on the vertices. Examples of these are the grayscale value of images, or height information in geographical data.

Let K be a simplicial complex with distinct real values specified at their vertices $h: K^0 \to \mathbb{R}$. We can then order the vertices by an increasing function value as $h(v_1) < \cdots < h(v_{m^0})$ where $m^0 = |K^0|$. Each simplex σ has a unique maximum vertex v_{max}, i.e.,

$$h(v_{max}) = \max\{h(v) : v \in K^0 \text{ and } v \prec \sigma\}$$

The *lower star filtration of h* [7, Section VI.3] is the nested sequence of complexes $\varnothing = K_0 \subset K_1 \subset \cdots \subset K_{m^0} = K$ such that:

$$K_t \backslash K_{t-1} = \{\sigma \in K : \text{maximum vertex of } \sigma \text{ is } v_t\}$$

In particular, K_t and K_{t+1} differ by at least one simplex since each simplex has a unique maximum vertex.

2.4. Persistent homology

Persistent homology [6,27] studies homology classes and their "lifetimes" (persistence) along a nested sequence of objects (simplicial complexes in our case).

Given a filter of K, the algorithm for computing persistence barcodes that appears in [6], marks an i-simplex σ_t as positive (birth) if it belongs to an i-cycle in $K_t = \{\sigma_1,\dots,\sigma_t\}$ (i.e., σ_t creates a new homology class at time t) and negative (death) otherwise (i.e., σ_t destroys a homology class created at some time s for $0 \leq s < t$).

Given a filter $\{\sigma_1,\dots,\sigma_m\}$, a *persistence barcode* [3] is a graphical representation of pairs of birth and death times, as a collection of horizontal line segments (*intervals*) in a plane. If a simplex σ_s creates a homology class at time s (the index in the filter) which is destroyed at time t, $0 \leq s < t \leq m$, then the interval $[s,t)$ is added to the corresponding persistence barcode. If a simplex σ_s, $0 \leq s \leq m$ creates a homology class at time s which survives along the process, then the interval $[s,\infty)$ is added to the persistence barcode. For a fixed i, the *i-barcode* is the set of intervals of a given persistence barcode corresponding to the pairs of positive i-simplices and negative $(i+1)$-simplices of K. See [3]. The following lemma holds.

Lemma 1. *Independent of the selected filter of a simplicial complex K, the number of intervals in an i-barcode,* $0 \leq i \leq n$, *is*

$$\beta^i + \sum_{j=i+1}^{n} (-1)^{i+1-j}(|K^j| - \beta^j),$$

[1] http://www.pub.ist.ac.at/courses/2012/computationalgeometryandtopology/Lectures/Lecture-11.pdf