UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

Dipartimento di Scienze
e Metodi dell'Ingegneria

Sede
Via Giovanni Amendola, 2
42122 - Reggio Emilia, Italia
T +39 0522 52 2161

www.unimore.it
www.dismi.unimore.it

Reggio Emilia, 26.1.2025

Review of the PhD thesis of Mr. Davide Gurnari

Dear Committee,

It is my pleasure to present the following review of the PhD thesis of Mr. Davide Gurnari entitled

"New Shape Descriptors for Topological Data Analysis".

1. Classification and contributions

The thesis under review is devoted to developing theoretical results and algorithms applicable to the topological analysis of data, with an eye on the need for parallelization when working with large datasets. Implementations are made available, and experiments are carried out on data from different fields: histologic images, knot polynomials, till gene and protein expression data. In particular, the focus is on the study of three popular tools in topological data analysis.

The first contribution consists of clarifying some properties of the Euler characteristic curves and providing distributed algorithms to compute them. It is carried out in terms of

- establishing a stability result: the L^1 distance between two Euler characteristic curves built on top of two filtrations of the same simplicial complex is bounded from above by twice the 1-Wasserstein distance between the corresponding persistence diagrams
- generalizing Euler characteristic curves from 1-parameter filtrations to multi-parameter filtrations in such a way that the above-mentioned stability property is preserved
- vectorizing Euler characteristic curves to input them into machine learning methods
- experimenting on cancer data

The main challenge with Euler characteristic curves is usually considered to be its lack of stability. The main selling point in this thesis is that this problem can be overcome by using L^1 metrics.

The second covered problem pertains to barcodes of persistent homology, where each bar corresponds to an equivalence class of cycles representing a topological feature. Harmonic representatives are known to exist and be unique. In this thesis, it is suggested that they can be used to build feature vectors for the data. This methodology is applied in this thesis for

- analyzing multi-omics data for cancer subtype prediction and unsupervised subtype detection

As a third topic in this thesis, the Mapper graph construction is broadened in scope and applicability by

- injecting more geometry into the data analysis requiring equivariance under a group action
- allowing to analyze maps between datasets

Chapters overview

Chapter 1 introduces the mathematical background for the thesis: simplicial and cubical complexes, homology and persistent homology. The material is comprehensive and gives a concise summary of the theory. It is generally well written but presents some typos and some oversimplifications. For example, in the definition of a persistence diagram, points are not confined to the half-plane above the diagonal; in the definition of the bottleneck distance, there is no mention of the cost of matching two points on the diagonal; same with the Wasserstein distance which is one of the main ingredients in the following chapter.

Chapter 2 introduces Euler characteristic curves and contains the main theoretical contribution of this thesis: the stability of these curves with respect to the 1-Wasserstein distance. I believe this result is correct although it should be underlined that, as Theorem 4 says, the 1-Wasserstein distance itself is not stable in the sense of Lipschitzianity but only of continuity as its upper bound depends on the number of simplices.

This theoretical part again contains typos and simplifications. For example, the name Fundamental Lemma of Persistent Homology is misused as the result here does not match the original. Also, the fact that the functions may be not summable is mentioned very late.

The algorithmic part of this chapter concerns the construction of a Vietoris-Rips complex based on a well-known structure, the simplex tree, but in a parallelizable way. Unfortunately, there is no review of the literature on this problem and no experimental comparison with other constructions. The pseudo-code is difficult to parse, in my opinion, because of what I believe are typos. However, I find interesting the observation about the effect of choosing a different order for the vertices.

I think the results of this chapter resonate with the pioneer work [Jonathan E. Taylor. Robert J. Adler. "Euler characteristics for Gaussian fields on manifolds." Ann. Probab. 31 (2) 533 - 563, April 2003] about the usage of the Euler characteristics in applications, which I am surprised there is no reference of. Moreover, works about the Euler Characteristic Transform are likewise not mentioned. A suggestion for future work is the application of the algorithm developed here to the Euler Characteristic Transform.

Chapter 3 reviews harmonic persistent homology from the literature and gives new insights into the properties of harmonic representatives. Then, it uses harmonic representatives for machine learning tasks in a way that is original, in my opinion. Also, I think that the conjectures about the harmonic cycle that emerged from the experiments deserve further investigation.

This chapter still suffers from typos but, once again, I believe the theoretical results are correct.

Chapter 4 introduces mapper-type algorithms. In particular, EqBM allows for building a mapper graph that reflects the dataset symmetries. This is obtained with data augmentation. Ball mapper instead is used to explore theorems about polynomial invariants in knot theory. Both contributions are very interesting and innovative.

Chapter 5 concludes the thesis by introducing ClusterGraphs, a construct that turns the output of a clustering algorithm into a graph to highlight interconnections among clusters. The visualized examples and the experiments are very convincing. IN the chapter there is also a quick mention to intra-connections within clusters, but this aspect is not developed. This also deserves further investigation in my opinion.

Unfortunately, the thesis lacks a discussion section and outlook for future work so it is difficult to imagine how Mr. Gurnari imagines the work should evolve. However, as above underlined, this thesis paves the way for developments in many directions.

3. Recommendation

Contents: This work makes important improvements in the applicability of the three main topological data analysis tools: persistence, mapper, and Euler curves. The theoretical, algorithmic, and experimental parts are very well-balanced and make the research agenda very clear.

Knowledge: Mr. Gurnari demonstrates a very good understanding of the mathematical aspects of topological data analysis and the ability to turn it into efficient algorithms, as well as the capacity to apply it to real-world problems in original ways.

Presentation: The thesis is well structured, and the line of thought is immediately understandable. What could be further improved is textual correctness with respect to some typos and simplifications, lack of running examples for the proposed algorithms, and at times comparison with the state of the art. However, it is certainly solid work.

Recommendation: In summary, the work is original, well-designed, comprehensive, and appropriately presented. Hence,

I recommend awarding the Ph.D. degree to the candidate

Cordially,

Claudia Landi

CLAUDIA LANDI
03.02.2025
07:50:09 UTC

Prof. Dr. Claudia Landi