Persistent Homology in Machine Learning: Applied Sciences Review

Oleksandr Yavorskyi[1], Andrii Asseko-Nkili[1] and Nataliia Kussul[1,2,3]
[1]*Department of Mathematical Modeling and Data Analysis, Igor Sikorsky Kyiv Polytechnic Institute, Peremohy Avenue 37, Kyiv, Ukraine*
[2]*Department of Space Information Technologies and System, Space Research Institute National Academy of Science of Ukraine an State Space Agency of Ukraine, Glushkov Avenue 40, Kyiv, Ukraine*
[3]*Anhalt University of Applied Sciences, Bernburger Str. 57, Köthen, Germany*
yaotianjiu@gmail.com, a0494034@gmail.com, nataliia.kussul@gmail.com

Keywords: Algebraic Topology, Persistent Homology, Machine Learning, Physics, Healthcare, Topological Data Analysis, Chemistry, Biology, Material Sciences, Data Processing.

Abstract: Topological Data Analysis ('TDA') has become a vibrant and quickly developing field in recent years, providing topology-enhanced data processing and Machine Learning ('ML') applications. Due to the novelty of the field, as well as the dissimilarity between the mathematics behind the classical ML and TDA, it might be complicated for a field newcomer to assess the feasibility of the approaches proposed by TDA and the relevancy of the possible applications. The current paper aims to provide an overview of the recent developments that relate to persistent homology, a part of the mathematical machinery behind the TDA, with a particular focus on applied sciences. We consider multiple areas, such as physics, healthcare, material sciences, and others, examining the recent developments in the field. The resulting summary of this paper could be used by field experts to expand their knowledge on recent persistent homology applications, while field newcomers could assess the applicability of this TDA approach for their research. We also point out some of the current restrictions on the use of persistent homology, as well as potential development trajectories that might be useful to the whole field.

1 INTRODUCTION

Artificial Intelligence ('AI') is a fruitful and flourishing area that focuses on the development of algorithms that are capable of replicating human behavior. An important constituent of this area, which encompasses a variety of mathematical instruments developed for capturing, formalization, and optimization of methods that can help in the aims of AI, called Machine Learning ('ML')

Linear algebra, statistics, and probability theory, as well as functional analysis, comprise the list of the most widely-used mathematical instruments for Machine Learning. At the same time, more sophisticated mathematical machinery receives ever-growing attention from ML specialists [1]. Algebraic topology could be considered as one of the most important of such 'mathematical newcomers' to the field. The impact of topology on the current ML scene led to the emergence of a whole new area called Topological Data Analysis.

Algebraic topology raises basic questions about the shape of the object and is especially interested in the shape features that are invariant under deformation. A hole in S^2 sphere or torus is a typical example of such an invariant since it does not diminish up until we 'cut' the figure.

2 FUNDAMENTALS OF PERSISTENT HOMOLOGY

Informally, persistent homology ('PH') allows us to discover which features of the data set (called 'point cloud') are time-invariant. The latter notion gives researchers the right means to assess some constant geometrical (or, it is better to say, 'topological') features of the set. Below we give a more elaborate introduction to this concept.

2.1 Complexes, Filtrations and Persistence

A simplicial complex could be defined simply as a set consisting of points, lines, and *n*-order polytopes of

some point cloud defined on a manifold. The dimensionality of the simplex is defined through its vertices, i.e., a *k*-simplex is represented by a convex hull with *k*+1 independent vertices [2].

Now, we could induce a family of simplicial complexes out of the point cloud *X*. This family should be ordered by increasing inclusion, such that complex K_n is included in K_{n+1} and so on. As an analog to this idea, one can refer to the notion of filtration used in measure theory and probability theory [3].

One now could ask a question on how long a given structure, formed by the simplicial complex at step *n*, is preserved in the dataset throughout the filtration. In this case, we can define the birth and the death time of a given feature and use this information to describe the dataset which is given. The persistent homology is then just a way to quantify the geometrical features that are preserved at step *n*. A visual explanation of these concepts can be found below in Figure 1 and Figure 2.

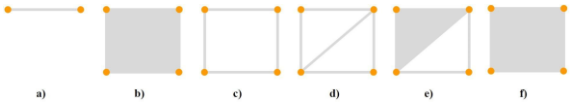

Figure 1: Different types of geometrical structures that possess different simplicial descriptions.

Figures a) and b) are the 1-simplex and 2-simplexes, respectively.

2.2 Persistent Homology as an Input

Before starting the discussion on the use of PH in Machine Learning, we should first consider how one could extract some meaningful information from it. By looking at the objects c) - f) in Figure 1, one could say that a 1-dim hole starts its existence at stage c) and lives up until stage e).

There are a number of ways to formalize and visualize this idea. Probably, one of the most common is the persistence diagrams akin to the one shown in Figure 2. On the diagram, each point represents a certain topological feature native to the point cloud analyzed (e.g., a hole). The coordinates on the XY plane of each point encode its birth and death times. Another popular way of showing this information is a barcode diagram, where instead of points, we use lines, and the length of each line denotes the lifetime of the feature. This information can be transformed [4] in order to get the persistence images, which can then be fed to the algorithm. Another option is to derive statistical information from the barcodes or diagrams, like the mean of the death time or average persistence length (which is *death time - birth time*).

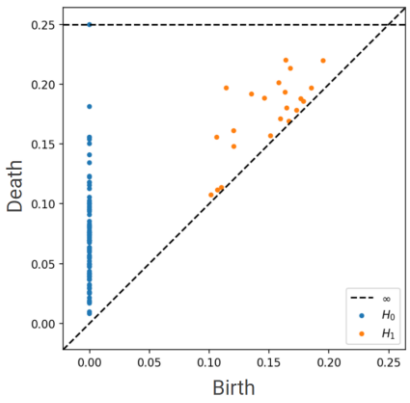

Figure 2: A persistence diagram of the random 2-dimensional point cloud. Features tracked in each dimension are shown in blue for H_0 and orange for H_1.

Normally, features that have a longer barcode (those located far from the diagonal on the diagram) describe features whose associated simplicial complexes are stable under various deformations (i.e., time-invariant from the filtration point of view). At the same time, those that die quickly tend to be treated as noise. The latter statement might be too expensive for areas where short-time transformations are important, such as chemistry, biology, drug design, physics, etc.

In this way, PH is able to provide two important insights about the data: does the data (as a point cloud) have a meaningful (useful) inner geometry and how to extract it; does the data (as a particular exemplar in the dataset) can be characterized based on its topological or geometrical features.

3 PERSISTENT HOMOLOGY IN MACHINE LEARNING

The opportunities for the use of PH information for Machine Learning lies in the concept of persistence of certain features of the dataset or its parts. This allows the characterization of the intrinsic geometrical features of the data available for the training. It becomes especially useful in cases when the geometry of the features might be put to use to crack the targeted problem. For example, in image recognition, where data is inherently geometrical [5-6].

At the same time, numerous data types can be transformed in a way that produces some unique geometry that corresponds to the data. The use of

Takens's theorem [7] to reconstruct the dynamical system behavior from the time series is one of the classic examples of such an approach.

Following this mindset, a great number of results were produced with the use of PH in Machine Learning for time series analysis [8-9], (computational) biology and chemistry [10-11], and even Natural Language Processing [12]. The general idea behind the scenes is to pre-process data using the PH-driven transformations and derive some characteristics of the objects (or the whole point cloud). The latter is then used for Machine Learning models. As a basic example, one can use the data on molecules or some materials' structure and apply PH analysis to obtain topological features of the molecules or materials and use this information for labels in classification.

A comparatively high area of applications of PH comes at a cost: most implementations are adjustments of some core ideas, which are hard to categorize and track; non-mathematicians might find it hard to cope with the nuances of these adjustments; the implementation of PH for Machine Learning requires a rather large amount of data preprocessing, which might create complications for experts outside of the industry. All these aspects related to the use of PH create a comparatively high 'entrance level', which leads to the low productivity of research. Below, we present an overview of the latest applications of persistent homology in Machine Learning in areas such as healthcare, physics, finance, and more. This coverage is intended to identify existing problems that are combated via PH, as well as the main techniques used during the problem-solving process.

3.2 Chemistry, Biology and Healthcare

This part is dedicated to the review of PH applications in Machine Learning problems that relate to Biology or Chemistry, including such practical applications as drug design and, generally, Healthcare.

As mentioned before, PH is a great tool when it comes to geometrically-enhanced data, such as images. A great example of how PH could be used for Magnetic Resonance Images ('MRI') analysis was proposed in [13]. The authors have proposed a DTA framework, which includes **D**ynamic Hierarchical Network Construction, Dynamic **T**opology Quantification, and Topological Pattern **A**nalysis. PH is used in multiple steps aiming to provide a topological description of the point cloud, which represents the MRI results. This description is then transformed and used along with a Balanced Random Forest (BRF) and Cost-effective Support Vector Machine (CE-SVM). The results of the detection of spatial patterns of multifocal lesions on clinical MRI were then compared to the existing approaches, showing a feasible increase in metrics. A great feature of the work is the authors' intention to additionally optimize their algorithm. The corresponding part might be used by other authors in their research.

Another Healthcare-related application can be found in [14]. Here, authors compare PH-induced metrics, such as (Slope of the) Betty Number Plot (BNP), BNP AUC, Mean of the area of persistent landscapes for components/holes, and others, with graph-induced metrics from the statistical point of view. In addition, the authors apply SVM for the classification comparison. The paper states that PH-induced metrics outperform their peers for the connectivity classification tasks, defined for Autism spectrum disorder conditions.

In [15], the authors use PH to compute coarsened topological features of atoms. After receiving the corresponding diagrams, they use Gaussian kernels for imagery transformation and then feed this information, along with other data, to a custom Neural Network in order to capture the protein dynamic information. The interested reader is strongly advised to look at the GitHub page provided by the authors to test the proposed approach.

As was shown in [16], functional cell description can be used in combination with PH to shed light on ionizing radiation-induced dysfunction in vascular endothelial cells. Akin to the approaches discussed before, the authors fuse the PH results with other data to get the molecular signatures of vascular dysfunction, which are then fed to a specific GANs-based algorithm to produce the results.

3.3 Physics and Geoscience

Persistence Homology receives ever-growing appreciation in material sciences due to its ability to work with shapes in a meaningful and deep way. At the same time, glass properties and structure is a long-lasting research field in material sciences, so there is no wonder that PH has found multiple applications in this area. A profound overview of this topic was published this year in [17].

Another research dedicated to the structural problem can be found in [18]. The paper discusses PH as a tool to analyze the structure of porous materials and Machine Learning approaches that can be used based on the results of PH in order to generate more efficient materials and model them in a more reasonable way.

A similar approach was followed in [19], here, the fuse of PH and Machine Learning is rather sequential than simultaneous. The paper discusses the filler morphology and how their properties can be analyzed using the combination of the aforementioned tools.

Developing the engineering trajectory of our discussion, an interesting paper covering slugging flow detection via the PH-induced techniques should be mentioned [20]. After Takens' theorem-based relevant time series transformation, authors use PH to track topological features of the obtained dynamical systems, This information is then used for the classification of the flow regimes.

An interesting example of how PH can be used even in the case of transfer learning is presented in [21]. In this paper, the authors examine the problem of chatter detection. In order to implement PH-induced metrics, authors have used a series of data transformation techniques, including FFT, which is a rare guest in PH and Machine Learning composition. The latter was presented by the use of SVM, logistic regression, random forests, and gradient boosting classification approaches.

In [22], authors implement another version of PH called zigzag persistence for crop insurance in agriculture. The key idea behind this concept is an ability to index the dynamics of the topological feature through their lifetime, providing a more comprehensive view of the point cloud structure. This approach was paired with the LSTM network, showing the reduction in the mean and variance of prediction error.

The sky is not the limit for PH applications. In [23], Large Scale Structures ('LSS') that refer to the patterns of galaxies or other objects of the universe were studied with the use of PH-related methods. More precisely, Copernicus Complexio's warm and cold dark matter models ('WDM' and 'CDM') were analyzed. It was shown that it is possible to develop a statistics-based approach, which uses the results obtained from the PH assessment, capable of distinguishing between the WDM and CDM. Moreover, the authors show that the scale at which differences occur is also trackable, which is especially important when dealing with LSS.

3.4 Selected Overview Papers

This section is intended to cover some important overview papers that came to light this year from each of the areas discussed above.

Chemistry experts can refer to [24] in order to examine the existing discussion on Materials Chemistry and how PH could be used for this purpose. Biology and drug discovery fields have recognized PH applications in [25-26], where a combination of Deep Learning and PH has become a major topic.

Healthcare topics were covered in [27-28], focusing on cancer detection, which is an important topic in AI-aided Healthcare and Precision Medicine.

4 DISCUSSION AND FUTURE WORKS

Persistent homology comprise a promising part of the Topological Data Analysis mainstream. As we have seen through the overview of the newly published papers dedicated to this approach, PH applications are mainly related to problems that have a geometrical nature. Biology and chemistry areas, including drug discovery as a joint field, tend to implement PH for cases where the problem is formulated around cellular, molecular or atomic structure, indicating the use of PH for mainly pre-processing purposes, deriving the characteristics of the objects being analyzed. Healthcare applications are, in turn, centred around image-driven data. PH methods tend to alleviate the same pain points, allowing a more meaningful approach for data extraction before the actual use of Machine Learning algorithms. At the same time, physics and material science enjoy PH in a similar way biology and chemistry do. These areas focus on geometrically-enhanced data to solve the emerging structural or dynamic problems implementing PH for data analysis as a part of the Machine Learning pipeline.

PH is rarely seen as a self-sufficient approach and tends to require an additional algorithmic pipeline. This implies an increasing production complexity for practical applications. Thus, a solution that automatically incorporates the PH-induced metrics and calculations would be highly appreciated by scientific and industry participants.

PH tends to be actively used for tasks that involve classification and seems to be flexible enough to be implemented with a variety of them. It's important to state that PH is used actively in combination with Deep Learning and generally Neural Networks-driven algorithms. Given an ever-increasing amount of data available, this is rather a good sign for PH, indicating a future development potential. At the same time, the scope of questions PH is capable of covering is not yet well-defined. For example, there is a limited conversation on how PH can be implemented for practical aspects of Meta Learning,

e.g., for synthetic data generation. It can be stated that expanded adoption of PH requires the development of new ways of data ' geometrization', combining statistics (flash and blood of Machine Learning) and geometry/topology together. At the same time, the current paper covers most of the PH applications published throughout the year 2022, which poses a question on certain applications asymmetry compared to other areas, e.g., recommendation systems for marketing or Natural Language Processing for chat-bots.

In addition, there is a rather limited (if any) amount of self-sustainable PH data analysis pipelines that are capable of providing data insights. This brings another question of whether this type of algorithm can be formulated in an efficient and useful way.

We thus consider further development of the applicability analysis of PH. We will especially focus on the applications that relate to image processing problems, as well as synthetic data generation tasks, which are of great importance in the case of imbalanced datasets for both structured and unstructured data cases.

5 CONCLUSION

In this paper we have discussed the recent applications of the persistent homology-driven analysis in the context of Machine Learning. The overview shows a wide range of applicability of such methods. At the same time, it can be suggested that the place of such methods is still comparatively narrow, being mostly a preprocessing technique. Interestingly, despite the existence of topology-inspired neural networks [29], none of the recent PH applications implements such architectural solutions. Clearly, as stated in the previous section, the ease of use of PH techniques can and should be developed further. Yet, the absence of such a combination cannot be solely derived from this fact, which raises another important question on the optimal fusion of topological preprocessing and architectural solutions.

The current paper is limited in details and comparison of non-topological and topological preprocessing techniques, which is a valuable research topic we are hoping to cover in the future. In addition, we sincerely encourage our readers to acquaint themselves with the list of references provided below in order to investigate in greater detail whether PH methods can be used for their current or future research.

REFERENCES

[1] F. Hensel, M. Moor, and B. Rieck, "A Survey of Topological Machine Learning Methods," Front. Artif. Intell., Sec. Machine Learning and Artificial Intelligence, May 2021.

[2] H. Edelsbrunner and J. Harer, "Computational Topology: An Introduction," 2010.

[3] A. Klenke, "Probability Theory," Berlin: Springer, 191 p., 2008.

[4] H. Adams et al., "Persistence Images: A Stable Vector Representation of Persistent Homology," Journal of Machine Learning Research, vol. 18, pp. 1-35, 2017.

[5] P. Frosini and C. Landi, "Persistent Betti numbers for a noise tolerant shape-based approach to image retrieval," Pattern Recognition Letters, vol. 34 pp. 863-872, 2013.

[6] G. Carlsson, T. Ishkhanov, V. Silva, and A. Zomorodian, "On the local behavior of spaces of natural images," International Journal of Computer Vision, vol. 76, pp. 1-12, 2008.

[7] F. Takens, "Detecting strange attractors in turbulence," Lecture Notes in Mathematics. pp. 366-381, 1981.

[8] A. Karan and A. Kaygun, "Time series classification via topological data analysis," Expert Systems with Applications, vol. 183, November 2021.

[9] S. Majumdar and A.K. Laha, "Clustering and classification of time series using topological data analysis with applications to finance," Expert Systems with Applications, vol. 162, December 2020.

[10] Z. Cang and G.W. Wei, "Integration of element specific persistent homology and machine learning for protein-ligand binding affinity prediction," International Journal for Numerical Methods in Biomedical Engineering, vol. 34, 2018.

[11] J. Townsend, C.P. Micucci, J.H. Hymel, V. Maroulas, and K.D. Vogiatzis "Representation of molecular structures with persistent homology for machine learning applications in chemistry," Nat. Commun., vol. 11, 2020.

[12] X. J. Zhu, "Persistent homology: An introduction and a new text representation for natural language processing," IJCAI, pp. 1953-1959, 2013.

[13] B.W. Xin, J. Huang, L. Zhang, and et al., "Dynamic topology analysis for spatial patterns of multifocal lesions on MRI," Medical Image Analysis, vol. 76, 2022.

[14] A.T. Jafadideh and B.M. Asl, "Topological analysis of brain dynamics in autism based on graph and persistent homology," Computers in Biology and Medicine, vol. 150, 2022.

[15] Y. Chiang, W.H. Hui, and S.W. Chang, "Encoding protein dynamic information in graph representation for functional residue identification," Cell Reports Physical Science, vol. 3, July 2022.

[16] I. Morilla and Ph. Chan, "Deep models of integrated multiscale molecular data decipher the endothelial cell response to ionizing radiation," iScience, vol. 25, January 2022.

[17] S.S. Sørensena, T. Du, C. Biscio, L. Fajstrup, and M.M. Smedskjaer, "Persistent homology: A tool to understand medium-range order glass structure,"

Journal of Non-Crystalline Solids: X, vol. 16, December 2022.

[18] D.P. Gao, J.H. Chen, Z.T. Dong, and H.W. Lin, "Connectivity-guaranteed porous synthesis in free form model by persistent homology," Computers & Graphics, vol. 106, pp. 33-44, 2022.

[19] T. Kojimaab, T. Washiob, S. Harab, and M. Koishia, "Search strategy for rare microstructure to optimize material properties of filled rubber using machine learning based simulation," Computational Materials Science, vol. 204, March 2022.

[20] S. Casolo, "Severe slugging flow identification from topological indicators," Digital Chemical Engineering, vol. 4, September 2022.

[21] M.C. Yesilli, F.A. Khasawneh, and B.P. Mann, "Transfer learning for autonomous chatter detection in machining," Journal of Manufacturing Processes, vol. 80, pp. 1-27, August 2022.

[22] T. Jiang, M. Huang, I. Segovia-Dominguez, N. Newlands, and Y.R. Gel, "Learning Space-Time Crop Yield Patterns with Zigzag Persistence-Based LSTM: Toward More Reliable Digital Agriculture Insurance," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, 2022.

[23] J. Cisewski-Kehe, B.T. Fasy, W. Hellwing, M.R. Lovell, P. Drozda, and M. Wu, "Differentiating small-scale subhalo distributions in CDM and WDM models using persistent homology," Phys. Rev. D, vol. 106, July 2022.

[24] D. Packwood et al., "Machine Learning in Materials Chemistry: An Invitation," Machine Learning with Applications, vol. 8, June 2022.

[25] Ch. Chinmayee, N.A. Murugan, and U.D. Priyakumar, "Structure-based drug repurposing: Traditional and advanced AI/ML-aided methods," Drug Discovery Today, vol. 27, pp. 1847-1861, July 2022.

[26] Y. Skaf and R. Laubenbacher, "Topological data analysis in biomedicine: A review," Journal of Biomedical Informatics, vol. 130, June 2022.

[27] C-E. Minciuna and et al., "The seen and the unseen: Molecular classification and image based-analysis of gastrointestinal cancers," Computational and Structural Biotechnology Journal, vol. 20, pp. 5065-5075, 2022.

[28] S. Prabhu, K. Prasad, A. Robels-Kelly, and X. Lu, "AI-based carcinoma detection and classification using histopathological images: A systematic review," Computers in Biology and Medicine, vol. 142, March 2022.

[29] C. Bodnar and et al., "Weisfeiler and Lehman Go Topological: Message Passing Simplicial Networks," Proceedings of the 38th International Conference on Machine Learning, PMLR, vol. 139, pp. 1-12, 2021.