International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 17 (2018) pp. 13268-13271
© Research India Publications. http://www.ripublication.com

Topological Data Analysis for Machine Learning Based on Kernels: A Survey

Edilberto Mejia-Ruda
Professor, Department of Biomedical Engineering, Militar Nueva Granada University, Bogotá, Colombia.

Robinson Jimenez-Moreno
Professor, Department of Mechatronics Engineering, Militar Nueva Granada University, Bogotá, Colombia.

Ruben Dario Hernandez
Professor, Department of Biomedical Engineering, Militar Nueva Granada University, Bogotá, Colombia.

Abstract. Dealing with large amount of data has become one of the most important issues to solve in Machine Learning (ML). In this way, extracting meaning from complex data is an interesting approach while dealing with large amount of information, very rich features (e.g. Genetic Data has more over than 500.000 features), or both. In this sense, Topological Data Analysis (TDA) Techniques appears to be useful for both, extracting condensed information about large amount of data by topological shape analysis. This paper is a brief introduction through a few selected aspects of Topological Data Analysis research, as a framework to improve and develop Machine Learning algorithms (MLAs). It is focused on present the major problems of ML and show the advantages of using a topological data analysis approach for dealing with large amount of data applications. It is presented more precisely, with machine learning techniques based on kernels due to the fact that it is already proof that TDA is useful for these learning approaches.

Keywords: Machine Learning, Topological Data Analysis, Big Data, Learning with kernels.

INTRODUCTION

Over the past few decades, Machine Learning (ML) has become an independent research discipline that has not only provided the necessary base for statistical computational principles of learning procedures, but also has developed various algorithms that are regularly used for text interpretation, pattern recognition, and a many other commercial purposes [5]. However, Machine Learning still has some issues, mostly in the amount of data that is needed for some learning tasks. Because of this, the capability of dealing with big data has become an interesting approach for researchers.

In this sense, a new set of techniques, Topological Data Analysis (TDA), based on geometry and algebraic topology, have been developing since 1991, with the work of Carlsson et al [20]. This techniques offers several advanteges for data dealing, extracting condensed information about large data sets. This advantages seems to be useful to solve some ML data problems related with scale invariance, richness, and consistency, e.g. the Kleinberg theorem [1] [6], that states that There are no clustering algorithms that satisfy scale invariance, richness, and consistency. However, it was not found yet a clearly TDA approach to solve this problems. In this sense, this paper presented a premiliminar state of the art on TDA as a framework to do ML focused on kernel based techniques that is already proved is a good approach for Machine Learning via TDA [7]-[16].

In this reasearch, the problem of learning and decision making is fundamental to understand the role that is going to take TDA in ML, as well as the remarks of TDA in data theory applications (DTA). In this sense, in section II first it is introduced a general framework of the major problems related with data analysis in ML. In section III it is introduced the remarks and fundamentals of TDA and it's applications. In section IV it is presented a survey about some of the approaches that have been made in TDA for ML based on kernels. In section V it was made a brief discussion of the information condensed in the above sections. And finally in section VI it is concluded about the advantages that might have TDA improving MLAs based on kernels.

MACHINE LEARNING AND DATA ANALYSIS

Machine learning is a paradigm that may refer to learning from past experience (which in this case is previous data) to improve future performance [5]. In this sense, learning becomes a data dealing problem while most ot the MLAs tries on giving meaning to data to make automacally decissions without any external assistance from human. This means, using Machine Learning, a researcher seek an approach through which the machine come up with its own solution based on the example or training data set provided to it initially.

Thereby, diferent algorithms are introduced for diferent types of machines and the decisions taken by them. Designing the algorithm and using it in most appropriate way is the real challenge for the developers and scientists [22]. In this way, several approaches have been developed, i.e. supervised learning, unsupervised learning, semi-supervised learning, reinforcement learning, and so on [5]. These approaches differs in the way they treat data, nonetheless, most of them use the concept of pattern recognition to make optimized decisions. As a consequence of this, according to Murphy [13], probability and statistics perspective seems to be the best way to tackle learning tasks. For example, in [4], Ghahramani gave a brief overview of unsupervised learning from the perspective of statistical modelling by using factor analysis, principal component analysis (PCA), mixtures of Gaussians, independent component analysis (ICA), hidden Markov models, state-space models, and many variants and extensions. He further concluded that statistics provides a coherent framework for learning from data and for reasoning under uncertainty. From this perspective, it should be noted, that ML has succeeded successfully significant real-world applications in several areas, such as Speech recognition, Computer vision, Bio-surveillance, Robot control, Accelerating empirical sciences, among other [9].

However, as stated by Talwar and Kumar [22], learning is a complex process as lot of decisions are made and also it depends from machine to machine and from algorithm to algorithm how to understands a particular problem and how to respond to it. Two of this major issues are related with noise (Data anomalies) and statistical variations related to data coordinates. As it is going to see later, TDA techniques does not present this issues, while they are inmune to noise and they are not altered by the data coordinates. As well as Talwar and Kumar, Mitchell [9] propose in a similar way some interisting research questions in the development of MLAs and their underlying theory that could be useful to solve some ML issues; How can we transfer what is learned for one task to improve learning in other related tasks? What is the relationship between diferent learning algorithms, and which should be used when? For learners that actively collect their own training data, what is the best strategy? To what degree can we have both data privacy and the benets of data mining? Can we build never-ending learners? can we develop a general theory of perception grounded in learning processes? as it can be noted, in someway, this questions expose the currently main problems for researchers in ML, and in this way it is expected that by looking learning paradigm from another perspective (e.g. TDA) some of this questions could be answered.

TOPOLOGICAL DATA ANALYSIS

According to Rucco et al [17] every moment of our daily life belongs to the new era of Big Data. We continuously produce, at an unpredictable rate, a huge amount of heterogeneous and distributed data. The classical techniques developed for knowledge discovery seem to be unsuitable for extracting information hidden in these volumes of data. Nevertheless, there is a set of algorthms based on algebraic topology, known as Topological Data analysis (TDA), focused on dealing with this large amounts of data. These techniques have been successfully used in diferent topics, according to Offroy and Duponchel [12], like gene expression profiling on breast tumors [11], T-cell reactivity to antigens for different type of diabetes [18], viral evolution [2], population activity in visual cortex [21] but also on unexpected topic as 22 years of voting behavior of the members of the US House of Representatives [8], characteristics of NBA basketball players via their performance statistics [8].

Topological Data Analysis was proposed first by Gunnar Carlsson [1] [20] and other mathematicians as a topological application to analyse datasets, given the unprecedented rate of data production. They found that Geometry and topology are very natural tools to treat much noiser data with more missing information, as biological data. They developed a mathematical formalism frame for incorporating geometric and topological techniques to deal with point clouds, i.e. finite sets of points equipped with a distance function. Carlsson explains the advantages of using this geometric methods by giving some key points and describing why topological methods are appropiate to dealing with them. Some of these are listed below as follows:

- ***Quantitative information is needed.*** Topology is exactly that branch of mathematics which deals with qualitative geometric information. This includes the study of what the connected components of a space are, but more generally it is the study of connectivity information, which includes the classification of loops and higher dimensional surfaces within the space.
- ***Metrics are not theoretically justified.*** Topology studies geometric properties in a way which is much less sensitive to the actual choice of metrics than straightforward geometric methods, which involve sensitive geometric properties such as curvature.
- ***Coordinates are not natural.*** Topology studies only properties of geometric objects which do not depend on the chosen coordinates, but rather on intrinsic geometric properties of the objects. As such, it is coordinate-free.

Fundamentals

In everyday applications, according to Patania, Vaccarino and Petri [15], there are two main TDA techniques: Topological Simplification (via the mapper algorithm [20]) and Persisten Homology [3] [23]. In this document, it is introduced the principles of persistent homology for single variable functions while it is the most widely used algorithm. For more information look at [10].

Let be a sample P:

$$X_1, X_2, \dots, X_n \sim P$$

Where P is supported on some set $X \subset R^d$ and X_i is a data point. X is noted to be a topological space.

Persistent Homology An intuitive idea of persistent holomogy is given in Fig. 1.

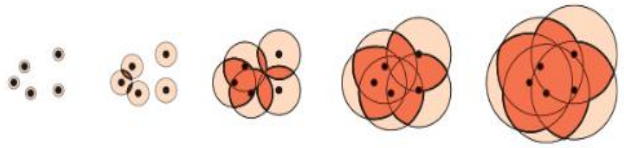

Figure 1. The $\in -$offsets. [19]

There, it can be seen a point cloud of data and the uninon set of balls $B(X_i; \in)$ for various values of _ described by:

$$\bigcup_{i=0}^{n} B(X_i; \in)$$

The key observation is the topological features appear and disappear as ∈ increases. TDA, uses the concept of homology to characterize sets based on connected components and holes using Betti Numbers, while holes description is a fundamental characteristic to do shape analysis using topology. In summary, persistence homology attempts to assign topological invariants to statistical data sets by using Betti Numbers descriptions. In this sense, they can be understood as fundamental characteristics for data analysis, because they are going to be used as the principal tool to analyse data. A very brief explanation of them is describe below.

Let $\beta_0, \beta_1, \dots, \beta_n$ be Betti Numbers, β_0 is number of connected components, β_1 is the number of one dimensional holes, β_i is the number of two-dimensional holes and son on. More formally J^{th} is the rank of the j th homology group. In general, Persistent homology examines these homological features from a multiscale perspective to analyze this datasets.

Topological Data Analysis for Machine Learning Based on Kernels

As it was already mentioned above, TDA has been sucessfully used for diferent topics, notwithstanding, it is not found yet enought research on TDA combined with ML techniques. On the other hand, it was found that must of research works are foused on doing ML based on kernels via TDA. Herewith is enclosed the clearest applications in this area, as researchers introduce TDA in a formal way to do ML based on kernels.

In [7], Kwitt et al uses an statistical treatment of persistence diagrams (a general approach that encodes persistent homology) to generate a universal persistence scale space kernel leveraging the theory of embedding probability measures into reproducing kernel Hilbert spaces. In this research, they proof that the functions of this kernel are contained on a Hilbert space. In this sense, this kernel is useful for Machine Learning based on kernels, as Support Vector Machines (SVM). Similarly Zhu et al [24] propose an stochastic multiresolution persistent homology kernel that is also useful in SVM applications. In this case, Zhue et al, proofs the advantages of using this kernel in some tasks as clustering and classification. Their kernel proved to be multiresolution and it can handle large point clouds.

In another work, Reininghaus, Bauer and Kwitt [16] presented a full integration of TDA for ML by determining a stable multi-scale Kernel for persistence diagrams. In this work they stablished a connection between this kernel and two kernel based learning techniques, SVM and PCA. Experiments on two benchmark datasets for 3D shape classification/retrieval and texture recognition show considerable performance gains of the proposed method compared to clasic Heat Kernel Signature (HKS) approach based on persistence landscapes. In summary, their method enables the use of topological information in all kernel-based machine learning methods.

In a much more general way, extending TDA kernel approaches for ML, in the Padellini and Brutti research [14] it is investigated the predictive power of TDA in the context of supervised learning by defining a topological exponential kernel that can be successfully used in regression and classification tasks. Results presented are encouraging for the emerging branch of TDA in general while they demonstrates practicality of this new set of tools for ML.

DISCUSSION

With this research it has been shown that TDA techniques could be an interesting approach for solving some data dealing problems in ML, as relating different learning algorithms [16] [14]. Results presented on [7] [24] [16] [14] demonstrates applicability of TDA techniques and proof they are useful to improve MLAs performance in terms of accuracy as is mentioned in [16]. On the other hand, it was not found enought research on this topics and it was observed that, as a pure mathematical field, researchers are mainly focused on proving mathematical properties of the developed techniques, e.g. [7].

CONCLUSIONS

As Padellini and Brutti mentioned [14], TDA is an exciting new field that has seen a tremendous growth in the last couple of years, however, theoretical developments have not been matched with popularity in applications, as topological summaries are defined in rather complex spaces. In the other hand, as it was mentioned in the section above, it can be concluded that, as a pure mathematical field, research contributions on ML based on kernels applications are not still enought to stablished a reliable set of techniques for MLAs improvement. However preliminary results obtained by researchers, shows that TDA could be a very good approach to solve some of the major problems of ML techniques based on kernels, as those described by Mitchell [9] in section II.

ACKNOWLEDGEMENT

The authors are grateful to the Nueva Granada Military University, which, through its Vice chancellor for research, finances the present project with code IMP-ING-2290 (2017-2018) and titled "Prototype of robot assistance for surgery", from which the present work is derived.

REFERENCES

[1] Carlsson G (2009) Topology and data, vol 46. DOI 10.1090/S0273-0979-09-01249-X, arXiv:1312.6184 v5

[2] Chan JM, Carlsson G, Rabadan R (2013) Topology of viral evolution. Proceedings of the National Academy of Sciences 110(46):18,566{18,571, DOI

10.1073/pnas.1313480110, URL http://www.pnas.org /cgi/doi/10.1073/pnas.1313480110

[3] Edelsbrunner H, Harer J (2008) Persistent Homology a Survey. Contemporary Mathematics 0000:1{26, DOI 10.1090/conm/453/08802

[4] Ghahramani Z (2004) Unsupervised Learning BT - Advanced Lectures on Machine Learning. Advanced Lectures on Machine Learning 3176(Chapter 5):72{112, DOI 10.1007/978-3-540-28650-9 5, URL http://link.springer.com/10.1007/978-3-540-28650-9fn g5fn%g5Cnpapers3://publication/doi/10.1007 /978-3-540-28650-9fn g5, 1512.00567

[5] Kajaree D, Behera R (2017) A Survey on Machine Learning: Concept, Algorithms and Applications. International Journal of Innovative Research in Computer and Communication Engineering 5(2):1302{1309, DOI 10.15680/IJIRCCE.2017.

[6] Kleinberg J (2002) An impossibility theorem for verisimilitude. NIPS pp 446{453

[7] Kwitt R, Huber S, Niethammer M, Lin W, Bauer U (2015) Statistical Topological Data Analysis - A Kernel Perspective. Advances in Neural Information Processing Systems pp 3070{3078, URL http://papers.nips.cc/paper/5887-statistical-topological-data-analysis-a-kernel-perspective

[8] Lum PY, Singh G, Lehman A, Ishkanov T, Vejdemo-Johansson M, Alagappan M, Carlsson J, Carlsson G (2013) Extracting insights from the shape of complex data using topology. Scienti_c Reports 3(1):1236, DOI

[9] 10.1038/srep01236, URL http://www.nature.com /articles/srep01236

[10] Mitchell TM (2006) The Discipline of Machine Learning. Machine Learning 17(July):1{7, DOI 10.1080/026404199365326, URL http://www-cgi.cs.cmu.edu/f_gtom/pubs/MachineLearningTR.pdf, 9605103

[11] Munkres JR (1984) Elements of Algebraic Topology

[12] Nicolau M, Levine AJ, Carlsson G (2011) Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. Proceedings of the National Academy of Sciences 108(17):7265{7270, DOI 10.1073/pnas.1102826108, URL http://www.pnas.org/cgi/doi/10.1073/pnas.1102826108, arXiv:1408.1149

[13] O_roy M, Duponchel L (2016) Topological data analysis: A promising big data exploration tool in biology, analytical chemistry and physical chemistry. Analytica Chimica Acta 910:1{11, DOI 10.1016/j.aca.2015.12.037, URL http://dx.doi.org /10.1016/j.aca.2015.12.037

[14] P Murphy K (1991) Machine Learning: A Probabilistic Perspective. DOI 10.1007/SpringerReference 35834, URL http://link.springer.com/chapter/10.1007/978-94-011-3532-0fn g2, 0-387-31073-8

[15] Padellini T, Brutti P (2017) Supervised Learning with Indefinite Topological Kernels pp 1{16, URL http://arxiv.org/abs/1709.07100, 1709.07100

[16] Patania A, Vaccarino F, Petri G (2017) Topological analysis of data. EPJ Data Science 6(1):0{6, DOI 10.1140/epjds/s13688-017-0104-x

[17] Reininghaus J, Huber S, Bauer U, Tu M, Kwitt R (2017) A Stable MultiScale Kernel for Topological Machine Learning 1412.6821

[18] Rucco M, Mamuye AL, Piangerelli M, Quadrini M, Tesei L, Merelli E (2016) Survey of TOPDRIM applications of topological data analysis. CEUR Workshop Proceedings 1748:1814

[19] Sarikonda G, Pettus J, Phatak S, Sachithanantham S, Miller JF, Wesley JD, Cadag E, Chae J, Ganesan L, Mallios R, Edelman S, Peters B, Von Herrath M (2014) CD8 T-cell reactivity to islet antigens is unique to type 1 while CD4 T-cell reactivity exists in both type 1 and type 2 diabetes. Journal of Autoimmunity 50:77{82, DOI 10.1016/j.jaut.2013.12.003, URL http://dx.doi.org/10.1016/j.jaut.2013.12.003

[20] Sheehy D (2014) (Multi)Filtering Noise for Geometric Persistent Homology pp 1-4

[21] Singh G (1991) Mapper : A Topological Mapping Tool for Point Cloud Data pp 182

[22] Singh G, Memoli F, Ishkhanov T (2008) Topological analysis of population activity in visual cortex. Journal of . 8(8):1{28, DOI 10.1167/8.8.11. Topological, URL http://jov.highwire.org /content/8/8/11.short

[23] Talwar A, Kumar Y (2013) Machine Learning: An arti_cial intelligence methodology. International Journal of Engineering and Computer Science 2(12):3400{3405, URL http://ijecs.in/issue/v2-i12/11ijecs.pdf

[24] Wasserman L (2016) Topological Data Analysis. ArXiv e-prints 1609.08227

[25] Zhu X, Vartanian A, Bansal M, Nguyen D, Brandl L (2016) Stochastic multiresolution persistent homology kernel. IJCAI International Joint Conference on Artificial Intelligence 2016-Janua:2449{2455