International Journal of Applied Engineering Research ISSN 0973-4562 Volume 14, Number 2 (2019) pp. 584-588

A survey of Topological Data Analysis Methods for Big Data in Healthcare Intelligence

Milan Joshi[1], Dhananjay Joshi[2]

[1*] *Dept. of Applied Mathematics, Mukesh Patel School of Technology Management and Engineering Shirpur, SVKM's NMIMS (Deemed to be University), Mumbai, India.*

[2] *Dept. of Computer Engineering, Mukesh Patel School of Technology Management and Engineering Shirpur, SVKM's NMIMS (Deemed to be University), Mumbai, India.*

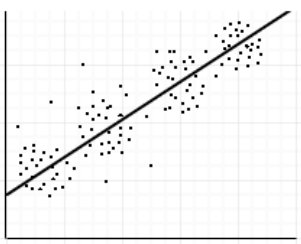
Abstract:

This paper provides an overview of topological methods inherited from Algebraic topology (computational topology) called Topological Data Analysis (TDA), to study Healthcare intelligence. In this work we will walk through two techniques from TDA namely Persistent Homology and Mapper. We will discuss how these techniques are effective, based upon the literature available where TDA has been applied in the context of Healthcare. This work is meant to direct future research efforts focusing on implementation of TDA in Healthcare intelligence for managing, discovering patterns and trends in healthcare data to do proper healthcare analysis for preventive and remedial decisions and automate healthcare.

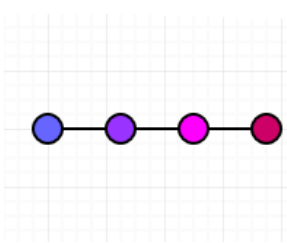
Keywords: Point Clouds, Topological Data Analysis, Network Graph, Persistent Homology, Mapper, Healthcare Intelligence, Machine Learning

1. INTRODUCTION

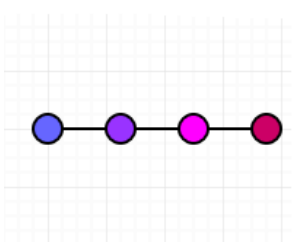
We hardly imagine life today without digital world. In the fields of medical, science, engineering, commerce, government, sports etc. large amounts of data sets are produced and processed daily. Such data requires analysis to extract meaningful and actionable insights, without which such data is waste. Since over the past few years, extraordinary efforts have been made by physicians and hospitals to adopt electronic health records, health system across the globe are creating large data warehouses and data takes to store these data on millions of patients and hundreds of millions of transactions per year. The data generated is messy, noisy, heterogeneous, high dimensional, dirty, sparse, unstructured, high dimensional, most important is humans (patients) are involved in healthcare data. Traditional analytics is slow, iterative process, have some missing insights and reliant on specialized individuals (Analyst and Data Scientist) asking rights questions of data and process it. But now we need to change the way we perceive the data, we need big ideas, different math, different algorithms to analyse and infer from such data. TDA is one such idea that can help identify patterns & provided useful insights, it has the ability to answer questions we didn't ask yet. We might not be nurses and doctors but the tools we create now and in the future are going to be directly influencing nurse and doctor decisions and we need to take that into account. As data engineers, data scientists and machine learning engineers, we have the ability to make tools to amplify the abilities of the medical professionals we support and we can make a huge impact. Application of topological data analysis and machine learning are growing together just in right time to address the healthcare, TDA is a new field of research it utilises topological concepts to classify and analyse data [1]. TDA is based on principle that data has shape and shape has meaning, meaning drives values [2]. The below diagram uses shape to analyse dataset using topology and the concept of shape where we see that we have four different clusters in fig 1 (a) running regression on it, fig 1 (b) and (c) is topological modelling of the same data. fig 1(b) using network graph models points on X axis where each node is one clusters and they are coloured according to low to high values (colour scale from blue to red). Similar for Y- Axis fig 1(c).



a: Linear Regression



b:Topological Modelling X Axis



c: Topological Modelling y Axis

Fig 1: TDA Summarizes the shape of data with no pre-convinced of what is should be

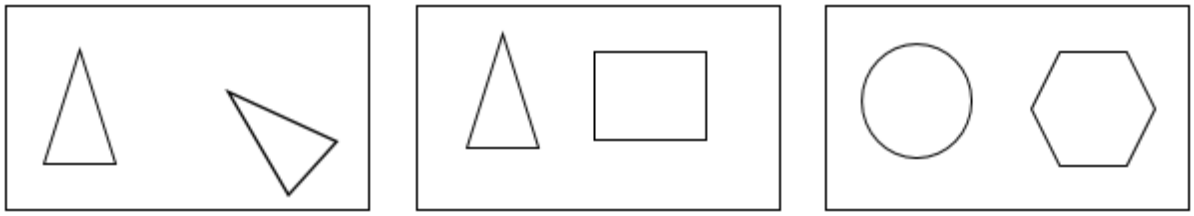
1.1 What is Topology?

Topology is a branch of pure mathematics, it is a part of geometry that focuses on the qualitative—not quantity but quality—in other words, the kind of position and the kind of underlying structure that defines the object. Topology is also called rubber sheet geometry. Any deformation, stretching, squeezing does not change topology. Gluing and tearing is not allowed, as it changes topology. If we take an object and stretch it like clay, relative position does not change. Thus, from a topological point of view, a cube and a sphere are equivalent. Famous example from topology where coffee cup is same as donut, shown in fig 2.

Fig 2: A coffee mug and a donut are equivalent in topology because they can be continuously deformed from one to the other [3]

Topology mainly deals with shape of data of data. The notion of shape is very fundamental one in the study of human perception. Shape display numerous wonders and highlights complex interconnections and intricate structures in data. Unlike traditional approaches TDA is Hypothesis free, no parameters with no choice of coordinates, TDA requires point clouds of data with pairwise distances without any scale, number of neighbours, noise bound.

TDA is 3-3-3 Model, 3 approaches (Homology, Mapper, Classification), 3 ideas or properties (Coordinate Invariance, Stretch or Deformation Invariance, Compressed Representation), and 3 benefits (Speed, Defensibility, Accuracy). In the figure below Coordinate invariance mean topology does not study properties of shapes which depend on the set of coordinates chosen. (studies done at different times, multi-variate data types collected) .Deformation invariance means when we stretch or deform geometric shape, we don't change the topological properties (real world data- human heterogeneity is complex and needs an approach that is deformation (variation) resistant) and compressed representation means any shape can be represented as network graph with finite points(this is so that signals can be easily identified in the form of "shapes" in the network) shown in the right most figure 3 that circle can be represented as hexagon sacrificing a little detail like curvature of circle .



Coordinate Invariance Stretch or Deformation Compressed Representation

Fig 3: Properties of topological analysis.

1.2 Topological Modelling of Data

There are two methods to model data using topology to find patterns and actionable insight from data in the first approach called persistent homology which tries to model the data by fitting topological space on the top of point cloud data with some distance (not necessarily Euclidean distance) defined on it. Persistent homology discerns the topological features of data. By "topological features," we mean things like components (or clusters) and holes in data. In this algorithm the persistence of every feature can be represented as a pair of two numbers in the form of interval (d_1, d_2) visualize as a bar from d_1 to d_2, where d_1 represents birth time of that feature and d_2 represent death time of that feature, short bars represent noise or artefacts in sampling long bar represents feature in data .

Fig 4: Persistence diagram and equivalent representation of barcode (right) for given sample data cloud. [4]

More detail discussion can be found in[5],[6],[7] .The next method we discuss is Mapper in this method First we project the dataset to low dimensions using Filter Functions You can use any projection function from maths, statistics, econometrics, or machine learning. Then we cover this projection with overlapping intervals/hypercube. Cluster the points inside an interval (either apply clustering on the projection and suffer projection loss, or cluster on the inverse image/original data). You can use any clustering algorithm

(hierarchical, density-based, etc.) and distance metric (does not need to be a proper metric that satisfies triangle inequality). The **clusters** become nodes in a graph. Due to the overlap, a single point can appear in multiple nodes. The more details on Mapper can be found in when there is such a member intersection, draw an edge between these nodes. The following figure shows the Mapper Procedure [8].

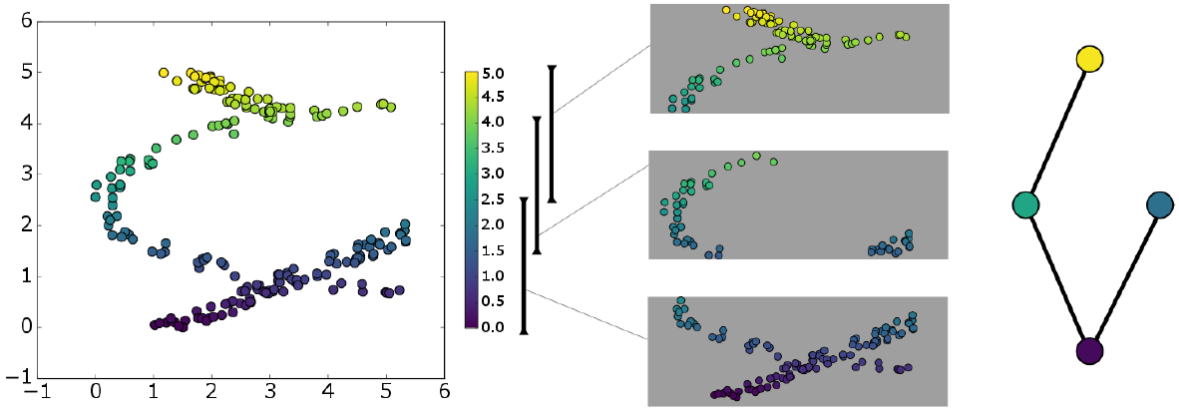


Fig 5: Mapper Procedure [8]

The following diagram shows the workflow of topological modelling of data for both the methods.

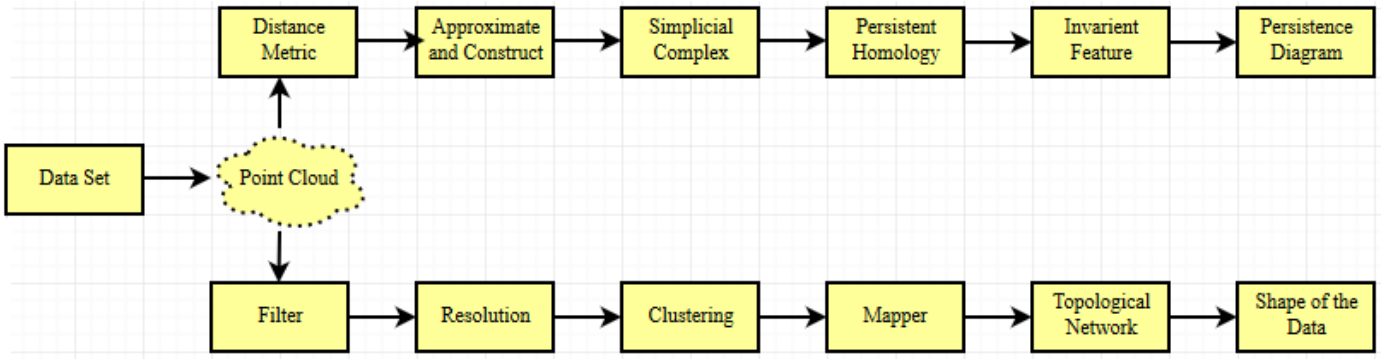


Fig 6: Steps involved in persistent homology and mapper.

1.3 Challenges in Healthcare

Actually there are lot of challenges in healthcare, but we will discuss the few here

Let's start our story with drug Failure. Eli Lilly and Company is a global pharmaceutical company headquartered in Indianapolis, Indiana. Had a drug failure in Phase III for Alzheimers Disease , Phase III means company have invested lot of money for the drug in for first II phases , why do such failure happened, one of the fundamental reason is actually limited understanding of human population (heterogeneous) and disease. Most of the things are carried out in clean environment where things are tested against animals. But in case of humans when disease is complex it's hard to find signal which make sense for our analysis, so need a way to deal with this situations.[9]

Another challenge is overcrowding or Excessive admissions in emergency departments. It causes problems for patients and staff which includes increased in waiting time, increased in length of stay, increased in patient mortality which causes lot of harm to hospital in terms of finance, and patient in terms of their health.

We will discussed how few of these challenges are addressed using topological data analysis methods.

2. LITERATURE REVIEW

2.1 Persistent Homology in Healthcare

TDA has the ability to detect the novel feature from data based on shape of data alone Persistent homology computes number of connected components, number of higher dimensional hole in data and represent those feature using persistent diagrams and barcodes [10].We see how these feature are detected in different TDA based data driven healthcare applications. In[11] author has studied Influenza net data, he uses persistent homology to study higher dimensional topological features and noise that are present in the data uses Persistent Homology in

TDA to analyse the topology of Influenza net data identifying noise and distinguishing higher dimension features. Further author compares the qualitative methods to other quantitative methods such as Fourier analysis or dynamical time wrapping (DTW)

As we have discussed in earlier section the problem of overcrowding in ED. In [12] author uses time series data and use persistent homology from TDA to propose two early real time indicators. One is for abnormal arrival at ED whereas second gives information on the time index of the maximum number of arrivals

For preventing and predicting overweight and obesity, author in [13] uses a variant of persistent homology called windowed approach to the persistent homology and combined with some other techniques. Based on time series data collected using tri-axial accelerometer using smart phone sensor author is able to detect the significant difference between two groups while other methods has failed to do so .This has significant implications for clinical decision making and patient care.

In [14] Author uses persistent homology and Machine learning combining some other methods other methods Image data for automating diagnosis of tumours to detect the tumors, which are classified into three classes: oval, tubular and irregular.

2.2 Mapper in Healthcare

We frequently discover subgroups and complex connections between subgroups in data, that conventional techniques are failed to find. TDA allows us to discover such subgroups with lot of subtle details and also helps in examining the complex interconnections between data sets. It is proven that TDA-based data-driven discovery has great potential application for decision-support for basic research and clinical problems such as outcome assessment, neurocritical care, treatment planning and rapid, precision-diagnosis. [15]. In Mapper we map the given finite point cloud to a combinatorial graph or network using filter function, number of bins and colours, further we can study the obtained network part be part, we can also do some statistical analysis on each nodes and mapper provides different subgroups of data that is useful to analyse, infer and detect novel features.

Type 2 diabetes T2D is complex and multifactorial disease that has emerged as an increasing prevalent worldwide health concern associated with high economic and physiological burdens using mapper algorithm from TDA Author has able to find different subgroups for T2D. Author hypothesize that a data-driven analysis of a clinical population could identify new T2D subtypes and factors. topology-based approach to (i) map the complexity of patient populations using clinical data from electronic medical records (EMRs) and (ii) identify new, emergent T2D patient subgroups with subtype-specific clinical and genetic characteristics. [16]. The following Analysis is carried out in Aysdi software by the author which is commercial leader and provide enterprise solutions using topological data analysis.

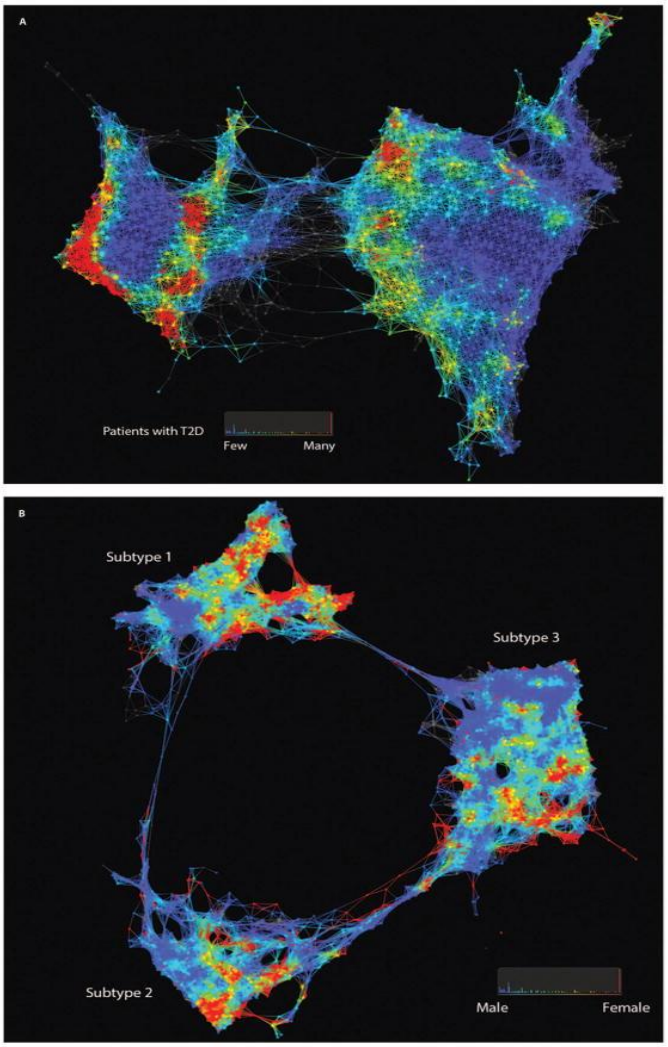


Fig 7: Patient and genotype networks [16]

The [17] author identified a unique subgroup of breast cancers. Which are known as c-MYB+ tumors with 100% survival rate.

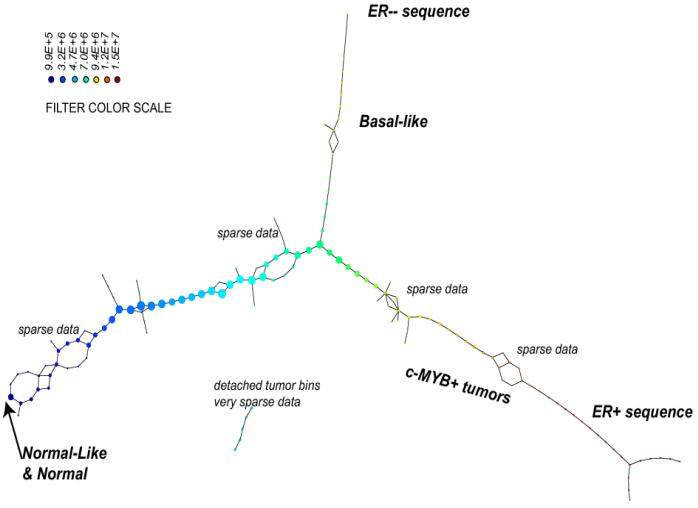


Fig 8: C-MYB+ Tumors subgroup with 100% survival rate obtained using TDA [17]

3. CONCLUSION

This survey mainly focused on performance enhancement of healthcare System. If appropriate data sets/images are available, Healthcare system will yield accurate result thereby facilitating the more correct diagnosis of the patient with proper preventive and remedial action. Proposed work will also give guarantee performances because of topological data analysis for machine learning, which is fast, data first approach, comprehensive and deep insights, more efficient, and more accurate. Machine Learning works magically with Topological data analysis.

REFERENCES

[1] Singh, G., Mémoli, F. and Carlsson, G.E., 2007, September. Topological methods for the analysis of high dimensional data sets and 3d object recognition. In *SPBG* (pp. 91-100).

[2] Carlsson, G., 2009. Topology and data. *Bulletin of the American Mathematical Society*, *46*(2), pp.255-308.

[3] https://www.cems.riken.jp/en/laboratory/qmtrt

[4] Gholizadeh, S., Seyeditabari, A., & Zadrozny, W. (2018). Topological Signature of 19th Century Novelists: Persistent Homology in Text Mining. *Big Data and Cognitive Computing*, *2*(4), 33.

[5] Ghrist, R., 2008. Barcodes: the persistent topology of data. *Bulletin of the American Mathematical Society*, *45*(1), pp.61-75.

[6] Edelsbrunner, H. and Morozov, D., 2013. Persistent homology: theory and practice.

[7] Fugacci, U., Scaramuccia, S., Iuricich, F. and De Floriani, L., 2016, October. Persistent homology: a step-by-step introduction for newcomers. In *Proceedings of the Conference on Smart Tools and Applications in Computer Graphics. Eurographics Association*.

[8] Munch, E. (2017). A User's Guide to Topological Data Analysis. *Journal of Learning Analytics*, *4*(2), 47-61.

[9] https://seekingalpha.com/article/4026439-eli-lilly-buy-recent-clinical-failure

[10] Zomorodian, A. and Carlsson, G., 2005. Computing persistent homology. *Discrete & Computational Geometry*, *33*(2), pp.249-274.

[11] Costa, J.P. and Škraba, P., 2015. A topological data analysis approach to the epidemiology of influenza. In *SIKDD15 Conference Proceedings*.

[12] Dugast, M., Bouleux, G., Mory, O. and Marcon, E., 2018. Improving Health Care Management through Persistent Homology of Time-Varying Variability of Emergency Department Patient Flow. *IEEE journal of biomedical and health informatics*.

[13] Biwer, C., Rothberg, A., IglayReger, H., Derksen, H., Burant, C.F. and Najarian, K., 2017. Windowed persistent homology: A topological signal processing algorithm applied to clinical obesity data. *PloS one*, *12*(5), p.e0177696.

[14] Dunaeva, O., Edelsbrunner, H., Lukyanov, A., Machin, M., Malkova, D., Kuvaev, R. and Kashin, S., 2016. The classification of endoscopy images with persistent homology. *Pattern Recognition Letters*, *83*, pp.13-22.

[15] Nielson, J.L., Paquette, J., Liu, A.W., Guandique, C.F., Tovar, C.A., Inoue, T., Irvine, K.A., Gensel, J.C., Kloke, J., Petrossian, T.C. and Lum, P.Y., 2015. Topological data analysis for discovery in preclinical spinal cord injury and traumatic brain injury. *Nature communications*, *6*, p.8581.

[16] Li, L., Cheng, W.Y., Glicksberg, B.S., Gottesman, O., Tamler, R., Chen, R., Bottinger, E.P. and Dudley, J.T., 2015. Identification of type 2 diabetes subgroups through topological analysis of patient similarity. *Science translational medicine*, *7*(311), pp.311ra174-311ra174.

[17] Nicolau, M., Levine, A.J. and 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*, p.201102826.