Topological Data Analysis of Two Cases: Text Classification and Business Customer Relationship Management

Bi Yue
Beijing-Dublin International College, Beijing University of Technology, Beijing, 100010, China
Corresponding E-mail: angela@cas-harbour.org

Abstract. Topological Data Analysis (TDA) is a rising and burgeoning field of data science, which provides a set of new topological and geometric tools to extract relevant features out of complex high-dimensional data. In this paper, we mainly study two papers: Topological Data Analysis of Time Series Data for B2B Customer Relationship Management [1]; An Introduction to a New Text Classification and Visualization for Natural Language Processing Using Topological Data Analysis [2]. We briefly introduce the topological concepts involved in the two cases, then compare and analyze the corresponding topological solutions. Accordingly, we summarize the advantages of TDA and finally draw a conclusion with some improvements to optimize them.

1. Introduction
Being a branch of data science, topological data analysis (TDA) is an emerging and fast growing field. As the field grows in popularity and new applications in marketing appear, TDA will get all areas of concern and acceptance. In this paper, two applications of TDA in business-to-business customer relationship management are discussed, which are applied in financial market and authorship attribution in the area of text classification. Researchers of the two applications did groundbreaking work in management and book sorting fields to explore TDA related solutions and broadened the development space of TDA. We study these two cases aiming to better learn and understand pioneered TDA methods by analyzing the applications of TDA.

2. Overview

2.1. Customer Relationship Management (CRM)
For a leading provider of cloud computing, a better understanding for the demand and expectation of its customer is required, thus improving its services. But the empirical data available at a customer level is limited and hard to work with. And popular heuristics such as the Recency, Frequency, Monetary (RFM) framework can be misleading according to figure 1 [1] that two customers can share the same RFM scores, but customer A is more alive than B. For the sake of improving the ineffective traditional forecasting techniques and generating reliable estimates for their future demand over multiple periods, the paper [1] proposes a new system to analyze the customer base (pre-processing customer data) and forecast demand in-depth. (Applied machine learning pipeline is depicted in figure 2.)
Figure 1. RFM framework [1].

Figure 2. Proposed machine learning pipeline with 3 different variations of the RFM framework [1].

2.2. New Text Classification and Visualization for Natural Language Processing
The rapid growth of Internet leads to explosion of information. How to get valuable information on the Internet rapidly and accurately is of great importance. We need to classify the text by assigning labels based on their content, this process is called text classification. Authorship attribution is one of the main branches in text classification which tries to identify the author of a given text based on its content. In the paper [2], the researchers proposes a new novel method other than semantic and non-semantic to do authorship attribution using Persistent Homology algorithm and Mapper with the application of authorship attribution of poems data set.

3. Topological solutions and analysis

3.1. Topological concepts involved
Simplicial complexes can be seen as higher dimensional generalization of graphs.

Given a set \( X = \{x_0, ..., x_k\} \subset \mathbb{R}^d \) of \( k+1 \) affinely independent points, the \( k \)-dimensional simplex \( \sigma = [x_0, ..., x_k] \) spanned by \( X \) is the convex hull of \( X \). The points of \( X \) are called the vertices of \( \sigma \) and the simplices spanned by the subsets of \( X \) are called the faces of \( \sigma \). A geometric simplicial complex \( K \) in \( \mathbb{R}^d \) is a collection of simplices such that:

- any face of a simplex of \( K \) is a simplex of \( K \),
- the intersection of any two simplices of \( K \) is either empty or a common-face of both [3].

Persistent homology tracks the changes of the homology of \( F \) as \( r \) increases, (a filtration \( \text{Filt} = (F_r)_{r \in T} \) of a simplicial complex or a topological space), like new connected components appearing, existing component merging, loops and cavities appearing or be filled, etc..., identifies the appearing features and associates a life time to them.

Changes mode example: Persistent homology thus registers \( a_1 \) as the birth time of a connected component and start to keep track of it by creating an interval starting at \( a_1 \). Then, \( F \) remains connected until \( r \) reaches the value \( a_2 \), where a second connected component appears. Persistent homology starts to keep track of this new connected component by creating a second interval starting at \( a_2 \). Persistent homology follows the rule that it is the most recently appeared component in the filtration that dies: the
interval started at $a_3$ is ended at $a_4$ and a first persistence interval is created, which encodes the lifespan of the component born at $a_3$... (according to figure 3.)

![Figure 3. The persistence barcode and the persistence diagram of a function $f: [0,1] \rightarrow \mathbb{R}$][2].

The persistence barcode encodes the span life of the different homological features encountered along the filtration using different format like: intervals, growing bars/balls etc., which are independent of the decomposition of persistence module $V$ [2]. (Persistence module: such a module can be interpreted as a feature that appears in the filtration at index $b$ and disappear at index $d$.)

A barcode represents each persistent generator (classes that generate $p$-th homology group) with a horizontal line beginning at the first filtration level, where it appears, and ending at the filtration level, where it disappears. In contrast, a persistence diagram plots points in $\mathbb{R}^2$ plane for each corresponding life-time interval with its $x$-coordinate referring to the birth time and its $y$-coordinate referring to the death time. A visual representation of persistent homology is persistent diagram.

The persistence landscape is a function $\lambda$:

- $N \times R \rightarrow R$, where $R$ denotes the extended real numbers $[-\infty, +\infty]$.
- It may be thought of a sequence of functions $\lambda_k: R \rightarrow R$, where $\lambda_k(t) = \lambda(k, t)$.
- Define $\lambda_k(t) = \sup \{m \geq 0\mid \beta^{l-m},(l+m) \geq k\}$ where $\beta_{i,j}$ is the dimension of group $\frac{H_i}{H_j}$. [2]

Persistent Homology Algorithm means, let $P$ be a point cloud data. First we construct the Vietoris Rips complex for $P$ as follows: consider an increasing sequence of positive real numbers $\epsilon_1 \leq \epsilon_2 \leq \epsilon_3 \ldots$, then we construct a cover of circles with centers of points in $P$ and diameter $1$, so we have as many circles as the number of data points in point cloud data, next we draw an edge between the center of each two circle which have any intersection and therefore we have a simplicial complex $VR(i)$. We do the same process for all $i = 1, 2, 3, \ldots$ as a result we have a filtration of complexes $VR(i)$. At last, we use the visualization method Barcode to analyze the information. Thus, a wide array of topological summaries from single complex objects can be computed directly in barcode.[2]

Mapper algorithm on knot shape data cloud:

- A data set $X$ with a metric or a dissimilarity measure between data points, a function $f: X \rightarrow R$ (or $R^d$), and a cover $U$ of $f(X)$.
- i.e. Project the whole data cloud to embedded space.
- For each $U \in U$, decompose $f^{-1}(U)$ into clusters.
- i.e. Partition the embedded space into overlapping bins.
- Then put data into overlapping bins.
- Next cluster the preimage $f^{-1}(U)$ of each open sets $U \in U$ using a cluster algorithm chosen by the user.
- i.e. use suitable clustering algorithm to cluster the points in the cloud data.
- each cluster of points in every bin would represented as a node of the graph and we draw and edge between two nodes if they share a common data point.
- At last, the output is the interactive graph.
When \( f \) is a real valued function, a standard choice is to take \( U \) to be a set of regularly spaced intervals of equal length \( r \geq 0 \) covering the set \( f(X) \). The real \( r \) is sometimes called the \textit{resolution} of the cover and the percentage \( g \) of overlap between two consecutive intervals is called the \textit{gain} of the cover.

\textbf{(NOTES: resolution: equal length \( r \) of each overlapping bins; gain: overlapping percentages of overlapping bins)}

The small changes in the resolution and gain parameters may lead to very large changes in the output of Mapper. The general strategy is exploring ranges of parameters and select the ones that turn out to provide the most informative output from the user perspective.

Let \( D_1, D_2 \) be two persistent diagrams and \( B \) be the set of all bijective functions \( \phi: D_1 \to D_2 \), then the \textit{Wasserstein distance} between two persistent diagrams \( D_1, D_2 \) denoted by \( W_p(D_1, D_2) \) is defined as follows. \cite{2}

\[
W_p(D_1, D_2) = \inf_{\phi \in B} \sum_{x \in D_1} \| x - \phi(x) \|_\infty^\frac{1}{p}
\]

(1)

\textbf{3.2. Case 1 solutions}

In the paper \cite{1}, the authors verify and compare two new methods: Time Series RFM and Topological applied to RFM bench-marked against the traditional RFM + Prediction to seek the effective method for higher performance of customer demand forecasting.

Firstly, in RFM + Prediction method, new attributes are obtained and they are embedded on each customer. They serve as new data features. And the enriched data set is then fitted in a appropriate predictive model to observe results.

Secondly, in this work, a time-series clustering technique based on shape” K-Shape” is used. The proposed method requires following phases: First, three time series were generated for each user corresponding to the Recency, Frequency and Monetary values. Thus, instead of having a point value for Recency, a time series is provided. Second, the prepared time series are used as an input in K-shape. Here, three instances are started. Each of them with four clusters. The number of clusters was decided through trial and error by visually inspecting the generated clusters. At last, the results from each of the three K-shape instances are embedded into the data. With the extended data set, a gradient tree boosting model is fitted.

Thirdly, \textit{topological RFM} is divided into five different steps. First, three time series are generated for Recency, Frequency and Monetary respectively. Then the time series are sliced using sliding windows. The objective is to generate delay embedding that can be projected as a point cloud. In addition, rips filtration is used to obtain the objective of obtaining death and birth complexes. Next, barcode diagrams are generated for both 0- and 1- dimensional homologies to visualize the birth-death filtered complexes. The focus is on the 1-dimensional homologies. Finally, a clustering is done through K-means based on features extracted from the barcodes. The number of clusters for each time series is decided using the Elbow method. With the obtained clusters, it is possible to enrich the original data set and use this information as additional features. \textit{Topological RFM} also uses gradient tree boosting for doing prediction.
Gradient tree boosting was used as a predictive model, the implementation catboost 7 was chosen and the obtained clusters were handled as categories. The data was divided into a training and test set using the 70-30 split. To compare the quality of the results, Root Mean Square Error (RMSE) was used. (Low values are better.) The results of the experiment can be found in figure 4.

### 3.3. Case 2 solutions

In the paper [2], researchers attempted and tested two approaches using Persistent Homology Algorithm and Mapper Algorithm to do authorship attribution for natural language processing, respectively.

In the experiment, researchers used the textual data (poems) of two Iranian poets Hafez and Ferdowsi. The data set gathered includes about 8000 different hemistich from both books. By applying Persistent Homology Algorithm, after preprocessing they fed the data to tf-idf algorithm in order to make document term matrix. Then they fed the document term matrix to persistent homology algorithm.

First, they sketched persistent diagram, barcode and persistent landscapes for a sample of Ferdowsi poems including 1000 hemistich. Furthermore, they also divided 8000 hemistich of hafez into 8 parts with the length of 1000, and computed persistent diagram and first landscape of each part. After that they sketched the mean landscape of these parts and did same work for 8000 hemistich of Ferdowsi. At last step, they computed Wasserstein distances between persistent Diagrams of correspondence parts of hafez and Ferdowsi poems. The computed results are shown in the figure 5 below.

![Figure 4. Overview of results using mean RMSE [1].](image)

![Figure 5. Computed results of Wasserstein distances between different parts of poems using Mapper Algorithm [2].](image)
By comparison, we can measure the minimum expected value of the distance between two different distributions utilizing Wasserstein distances, including the distance between discrete distribution and continuous distribution. Therefore, the results can be easily obtained (corresponding persistent diagram, Barcode and persistent landscapes are omitted).

Using Mapper method, researchers examined two accuracy tests on the shape graph. First, they partitioned the whole graph into 3 clusters (“Hafez”, “Ferdowsi”, “Both”), each named cluster include the high percent of corresponding poems (“Both” cluster have the same amount of both poems). The goal was to examine the percent of poems in each cluster really belongs to corresponding poems. Each cluster were processed through the following calculations:

Model preparing:
- Using “truncatedSVD” and “t-SNE” as filter functions after applying “TF-IDF” on the data.
- Choosing appropriate resolution and overlap: after comparing different resolution and overlapping amounts, they found that “the more resolution we have, the better data will be partitioned and classified and higher the overlapping percentage is, the more compact our resulting graph would be.”
- For clustering they used “Agglomerative Clustering” available in “sklearn” package with “cosine” similarity and complete linkage.

TDA Mapper in text classification STEPS:
- Gathering and pre-processing the data of poems.
- Using TF-IDF algorithm to turn pre-processed data into a matrix which columns represent a single word in the entire corpus and each row represent a hemistich.
- Zoomed data [100*100].
- Filtration step, Using truncated SVD and T-SNE algorithms as filter-functions.
- Bining: the whole 2-dimensional space divided in to smaller bins.
- Clustering: clustering algorithms applied to each bin in order to create the nodes in the final graph.

Figure 6. Experiment steps of TDA Mapper in text classification of the Hafez cluster [2].
Output of the algorithm as a graph, every node in the graph is a cluster of the previous step and nodes share an edge if they have a data point in common. (in the figure 6)

- calculation formula

\[
\text{Number of poems of this poet in a node of the cluster} = \frac{\text{Number of all poems in the whole cluster}}{\text{Number of all poems}}
\]  

So if we have the accuracy percentage of a for a cluster it means that a percent of the poems in that cluster has been labeled correctly.

After examining the first test on each cluster, we got the following results: for Hafez cluster percentage of accuracy was 80 percent, for Ferdowsi cluster percentage of accuracy was about 94 percent, and for Both cluster percentage of accuracy was 40 percent for Hafez poems and 60 percent for Ferdowsi poems. So for the Hafez cluster we can say that 80 percent of the poems in the cluster has the right label and so on for other clusters.

Second, they tried to divide some parts of the graph into several clusters based on their semantic subjects. To visually analyze each cluster were presented the text with in each cluster as a word cloud.

They can choose some clusters from the graph shape after the second test and then examined if they are semantically related to each other. (as figure 7)

4. Advantages and improvements

In the work [1], after constructing a time series, from a qualitative perspective, the benefit of using them is their highly visual component. The marketing analyst can show the results to decisions-makers. From a quantitative side, Topological RFM approach can be used as a clustering method to increase the accuracy of a predictive model for loyalty scoring. And the available data is easier to work with. Another benefit of using Topological RFM is the improvement on accuracy of a machine learning model. Thus, they open the door for a CRM predictive pipeline, where the practitioner can segment the customer base, generate personas, do predictions and communicate to manage both of the predictions for the personas as well as for the individual users and identify those users diverging from the expected results from their respective personas. Therefore, the provider can obtain a deeper understanding of its customer base to adjust its product offering and promotions while being able to generate reliable estimates for their future demand over multiple periods. Moreover, there are also some improvements that can bring more specific and accurate results for the model.

First of all, after constant demonstrations and a plentiful of experiments of value, we can also analyze differences between disparate public and proprietary data sets of data and identify the data set type
applicable to TDA from a theoretical perspective and from a practical one. Also, It is important to generate heuristics that can guide quantitative analysts in their choice.

In addition, for the next step, we should attempt to implement consensus clustering or clustering ensemble in order to avoid handling the three clusters of RFM as categorical variables, but rather merge them into a super cluster.

Furthermore, we can expand this work to very large data sets. For example, the work of [6] builds a tractable framework combining TDA with Optimal Transport to improve the efficiency of large-scale standard computing tasks of persistence diagrams.

5. Conclusion

As the first work dedicated to applying TDA techniques on CRM data to evaluate customer loyalty [1] and using topological models on poems to attribute authorship for NLP [2], the researches break new ground and verify these pioneering and successful approaches.

Moreover, there are many path-breaking attempts putting TDA into use, like the exploration of applying TDA to detect early warning signals of imminent market crashes [4]. And the introduced new time-series topology data-set, TS-TOP and a computational framework are applied for exploring the topology of time-delay embedding and processing and analyzing time-series data [5]. These researches expose new opportunities for applying TDA to new domains as well as increase the accessibility of relevant data-sets and advanced topological techniques.

As researchers continuing to innovate and applying TDA in various fields in different ways, it is to be expected that TDA will become an indispensable tool for marketing practitioners, managers, analysts, forecasters, research and development people and so on. The application of TDA can improve the performance of the work and work efficiency, so as to benefit the society development in all industries. We can predict a very positive outlook for the future of topological data analysis.

References

[1] Rivera-Castro, R., Pilyugina, P., Pletnev, A., Maksimov, I., Wyz, W. and Burnaev, E. (2019) Topological Data Analysis of Time Series Data for B2B Customer Relationship Management. https://arxiv.org/abs/1906.03956.

[2] Elyasi, N. and Moghadam, M. H. (2019) An Introduction to a New Text Classification and Visualization for Natural Language Processing Using Topological Data Analysis. https://arxiv.org/abs/1906.01726.

[3] Chazal, F. and Michel, B. (2017) An introduction to Topological Data Analysis: fundamental and practical aspects for data scientists. https://arxiv.org/abs/1710.04019

[4] M Gidea, M. and Katz, K. (2017) Topological data analysis of financial time series: Landscapes of crashes. https://arxiv.org/pdf/1805.08331.pdf

[5] Seversky, L. M., Davis, S. and Berger, M. (2016) On Time-Series topological data analysis: New data and opportunities. https://ieeexplore.ieee.org/document/7789621

[6] Lacombe, T., Cuturi, M. and Oudot, S. (2018) Large scale computation of means and clusters for persistence diagrams using optimal transport. https://arxiv.org/pdf/1805.08331.pdf