Topological Data Analysis of Customer Relationship

Yutao Lu
School of Statistics and Mathematics, Zhejiang Gongshang University, China
angela@cas-harbour.org

Abstract: In the modern society, if manufacturers want to make huge profits, they must not only control the production cost, but also classify customers and increase the loyalty of key customers. Thus, existing data analysis methods, represented by Time Series Analysis and Topological Data Analysis (TDA), have emerged at this point of time to explore customer relationships. Based on the latest research of time series analysis and TDA, application of analysis is explored in practice. The relationship between topology and datasets will be given by displaying the process of TDA. In the end, there will be a discussion about the efficiency of different approaches, the improvement of tools in this paper and their extended application. This work is intended to provide a brief introduction of TDA and other methods to help readers understand concepts easily and provide suggestions for future studies.

1. Introduction
In this era of rapid development, a large amount of various data is growing at an unpredictable speed every day, far exceeding the processing capacity of existing data analysis methods and hardware devices, and the result is called “Data Disaster”.

In order to improve the processing speed of data analysis, a great number of new methods have emerged in recent years. Among all the original methods, Time Series Analysis is a method that processes large and complicated datasets quickly, and Topological Data Analysis, [2] represented by simplicial complex, is an efficiency-improvement approach that conducts classification or clustering according to the shape of data or data’s topological structure.

This paper focuses on the issue that it’s hard to measure and classify the value of customers. Because of novelty and flexibility of the cloud datasets, and the historic data is limited, seasonality will be hard to detect and historic records are often not representative.

Besides, some common information variables, such as Recency, Frequency and Monetary [1] framework may be misled by some special situations. Meanwhile, commodity suppliers are interested in the deep demands of customers, in order to make reliable estimate of offerings in the future. Thus, the traditional FRM methods do not apply to the estimation of customer value, and it will be replaced by three different methods in this paper.

2. Basic Theories and Approaches

2.1 Basic Theories

2.1.1 Recency Frequency Monetary
Recency Frequency Monetary [3] originates from the marketing management process, which can be regarded as indexes of sales management. The RFM model is an important tool and means to measure
customer value and customer profitability. The above three indexes will be subdivided into five dimensions, so as to subdivide 125 types of customers. Then a 3-digit number will be obtained to classify the type of customers. In the end, according to each type of customers, goods and service providers can conduct precise marketing.

2.1.2 Time Series Analysis
Time Series Analysis is a data analysis method that divides a data series into four parts—Trend, Cycle, Period and Random factors, and then combine these parts to make future prediction. It emphasizes that continuous remote sensing observation of a region within a certain period of time is used to extract image features and analyze their changing processes and development scale [4].

The data of time series has two characteristics: 1) it is a real set of data, rather than a set from the mathematical experiments, which means that they reflect some phenomena in the type of index. 2) It is the data of dynamic environments, rather than stationary and static environments.

2.1.3 Topological Data Analysis
Topological data analysis (TDA) is one of the newest and fastest-growing methods in data science, analyzing data by focusing on the shape of the data and reducing the dimensions of the data. TDA is based on the methodology of Statistic and Algebraic Topology, and due to its methodology, TDA can also be a clustering method that is robust to noise. The goal of TDA is to reduce the dimensionality of high-dimensional data, analyze the topological structure or shape of data, and finally cluster the complex data.

2.2 Approaches

2.2.1 RFM + Prediction
With this method, three attributions can be obtained and inserted into the information of customer, which will serve as new features of the customer. The above process is the first step of data processing in machine learning, then the processed data will be fitted into the prediction model. In the paper, the gradient tree boosting is chosen because of its universality and advancement. The reason for using this method is to compare with the rest two methods as a control group.

2.2.2 Time Series RFM
In this paper, the shape-based time series technique “K-shape” is accepted. And K-means method is also included in this technique. There are three steps in this method: 1) several time series should be generated, representing the indexes of Recency, Frequency and Monetary. 2) The time series will be used as input in K-shape, and each will have four clusters. The number of clusters was decided by trial and error and visually inspecting the generated clusters. 3) The results of several instances will be embedded into the data, and then a gradient tree boosting can be constructed with such an extended dataset [5].

2.2.3 Topological RFM
In order to combine topology with RFM, five steps are required: 1) three time series of Recency, Frequency and Monetary will be generated, similar to the first step in Time Series RFM. 2) Use sliding windows to slice the time series. The objective is to generate a delayed embedding that can be projected as a point cloud. 3) After obtaining the point cloud, a popular algorithm, Rips filtration, will be used to obtain the death and birth complex. 4) Generate barcode diagrams for both 0-dimensional and 1-dimensional homologies. They help to visualize the birth-death filtered complexes. The key lies in the 1-dimensional homologies. 5) Finally, clustering should be performed with the method of K-means based on the features extracted from the barcodes. Elbow method can be used to estimate the number of clusters, which can enrich the original dataset by using the extracted information as additional features. Moreover, Topological RFM can be predicted by gradient tree boosting [6].
3. Review

3.1 Topological Concepts

3.1.1 Simplicial Complex
A simplicial complex [7] is a collection $L$ of finite non-empty sets, so if $K$ is an element of $L$, then is also true every non-empty subset of $K$.

Figure 1. Example of Simplicial Complex

The collection of 4 points, 3 lines, 1 triangle forms a simplicial complex.

Figure 2. Example of Non Simplicial Complex

The collection of 3 points, 3 lines, 1 triangle can not form a simplicial complex.

3.1.2 Simplicial Homology
Given $n \in \mathbb{Z}_+$, the $n$-th homology group of a simplicial complex $L$, is denoted by $H_n(L, \mathbb{R})$, and is defined as

$$H_n(L, \mathbb{R}) = \frac{Z_n(L, \mathbb{R})}{B_n(L, \mathbb{R})}$$

Where, $Z_n(L, \mathbb{R})$ is the kernel of the map $\partial_n$, which is the set of n-cycles. $B_n(L, \mathbb{R})$ is the image of the map $\partial_{n+1}$, which is the set of n-boundaries. $H_n(L, \mathbb{R})$ is a quotient vector space and the elements of $H_n(L, \mathbb{R})$ are equivalence classes of n-cycles of $C_n(L, \mathbb{R})$, which is a chain complex.

3.1.3 Betti Number
Given $n \in \mathbb{Z}_+$, the $n$-th Betti number [8] of a simplicial complex $L$ is denoted by $\beta_n(L)$, and is defined as

$$\beta_n(L) = \dim(H_n(L, \mathbb{R}))$$

This is the extended concept derived from simplicial homology.
3.1.4 Persistent Homology

Persistent Homology [9] is a widely used tool for data analysis, which is a method to solve the problem that whether the feature of data is persistent or not. Given the parameterized set in a certain space, this paper focuses on the invariant topological features with a large variation in parameters. The evaluation of persistence through the function \( f(x) \), where the function value \( f \) is parameter and the object is \( R_{f \leq t} \), is as follows.

Give the function \( y = f(x) \) as shown above.

When \( y \leq 1 \), the value of \( x \) is the collection of the left green bar, \( x = 2 \).
When $y \leq 4$, the value of $x$ is from negative infinite to positive infinite.

From the above figure, it is obvious that when $x = 1, 2, 3, 4$, the value of points can be a local maximum or a local minimum, which can be called a “critical point”. In general, when a parameter is changed, as long as it does not pass the critical point, the topological feature will not change.

3.2 Relationship between Topology and Dataset

The relationship between topology and Dataset can be represented by the process of TDA [10]. And the process is as follows:

1. Form an original point cloud from the dataset.
2. Color by filter value. Every point can calculate a filtered value by using a filter function, which can be a linear projection of the data matrix, a density estimation of the distance matrix, or a center degree index.
3. Divide by filter value. According to its filter value, the data points can be divided into different filtering value intervals, which are ranked from small to large. Besides, it’s important to note that the adjacent filter value intervals can be set with a certain overlap area, which means that the points in the overlap area belong to both intervals.
4. Cluster and network construction. After the last step, the data in each interval can be clustered separately. Then the small classes obtained from the previous interval clusters should be put together, and each small class can be represented by a circle of a different size. If there are identical original data points between the two classes (which is why the intervals must overlap each other), then an edge should
be added between them.

Here is a figure that shows the brief process of TDA:

A Original Point Cloud

B Coloring by filter value

C Binning by filter value

D Clustering and network construction

Figure 7. Brief Process of TDA [10]

4. Discussion

4.1 Efficiency of the Approaches

In this paper, there are three methods that are driven by the idea of RFM, that is, customer relationships evolve as the application time elapses, thus, the time series can be constructed.

From a qualitative perspective, the benefit of using them is that they have a high visual component. For instance, marketing analysts can continue to use their familiar tools to extend the concept and show
the results to decision-makers from the centroid, which is a time series as well and can provide important information about cluster behavior. This significant information can be delivered to the organization, which can help them in the creation of customer roles.

From a quantitative perspective, using Time Series RFM or Topological RFM can improve the accuracy of the model. Interestingly, compared with not using RFM, using RFM shows only minor improvements. But after adding Time Series method and Topological method, the accuracy of the model can improve a lot in machine learning. Besides, these methods can be used for future prediction, which can identify deviation from expectation value of customers, regardless of individual users or groups.

4.2 Improvement of the Tool
For the further improvement of the tool, based on the experiment in the paper, this work will seek to implement consensus clustering or clustering ensemble to avoid processing the three clusters of RFM as categorical variables, but rather merge them into a super cluster.

Another important step is to identify the sort of approaches from both theoretical and practical perspectives, using rather Topological RFM than Times Series RFM is a better choice. It is important to provide a new combination of approaches that can provide new choices to analysts. Another direction for improvement is to extend the range of application on big data, which is a huge demand in current society.

4.3 Application Extension
Besides the application for identifying the relationship of customers, current data analysis approaches, represented by Time Series Analysis and TDA, can be used in a variety of fields as well, as long as there are demands of exploring the information behind the data. For instance, these methods can be used in machine learning, for the approaches can study and build information structures by itself. And if someone leads the research, these methods can construct more complicated relationships.

In addition, TDA can also be used in the construction, physical and medical fields as well. In these fields, analysts can build the topological structure of data and analyze them to extract features, which may provide some significant information, while other methods cannot offer such special information.

5. Conclusion
Based on experimental research and the latest achievements of scholars, this paper explores the application of topological data analysis in practice, especially in classification of customer value and the handling of customer relationships. It finds out that time series analysis and TDA have high visibility and have tangibly improved the accuracy of prediction.

Although this paper has explored the application of time series analysis and TDA in RFM management, these models are only basic models, not complex models, and in this paper, only customer relationships are explored. Thus, a wide range of complex models and theories can be used in various research fields in future studies, such as using TDA in machine learning to help the AI systems develop their own way of studying and analyzing, or establishing topological structures to extract features in medical fields.

References
[1] Rodrigo Rivera-Castro, Polina Pilyugina, Alexander Pletnev and Evgeny Burnaev. Ivan Maksimov, Skolkovo Institute of Science and Technology; Wanyi Wyz, Alibaba Cloud Intelligence Business Group. Topological Data Analysis of Time Series Data for B2B Customer Relationship Management. arXiv:1906.03956v2 [cs.LG], Jun 17, 2019
[2] Carlsson, G. Topology and data. Bull. Amer. Math. Soc, 46, 255308 (2009).
[3] Sunil Gupta, Dominique Hanssens, Bruce Hardie, Wiliam Kahn, V. Kumar, Nathaniel Lin, Nalini Ravishanker, and S. Sriram. Modeling customer lifetime value. Journal of Service Research, 9(2):139–155, 2006.
[4] Saeed Aghabozorgi, Ali Seyed Shirinkhorshidi, and Teh Ying Wah. Time-series clustering - A decade
review. Information Systems, 53:16–38, 2015.
[5] John Paparrizos and Luis Gravano. Fast and accurate Time-Series clustering. ACM Transactions on Database Systems (TODS), 42(2):8, June 2017.
[6] Katharine Turner and Gard Spreemann. Same but different: distance correlations between topological summaries. March 2019.
[7] J.R. Munkres. Elements Of Algebraic Topology. Avalon Publishing, 1996.
[8] Frederic Chazal and Bertrand Michel. An introduction to topological data analysis: fundamental and practical aspects for data scientists. 2017.
[9] Hong feng. Fast calculation method of simplex coherence and its application in mass data analysis. National university of defense science and technology, 2011.
[10] Xiaodong Yang. Jian Shu. https://www.jianshu.com/p/81ab08c6f6d83, 2017.