Topological Data Analysis and Its Application to Time-Series Data Analysis

● Yuhei Umeda ● Junji Kaneko ● Hideyuki Kikuchi

The commercialization of AI technology has accelerated in recent years, with a growing interest in various machine-learning technologies such as deep learning. However, machine learning is based on statistical data analysis, and it is known today that certain information contained in such data is lost through analytical processes. To make the most of such information, we have developed a new machine learning technology based on topological data analysis (TDA) that focuses on and analyses the "shapes of data." This paper explains TDA as a new data-analytical method. As applied cases of TDA, it also describes the time-series deep learning for analyzing time series data and anomaly-detection technology, with an account of a bridge deterioration assessment in which the latter was applied.

1. Introduction

The recent appearance of "deep learning" is driving a third-generation AI boom that is now making inroads into society. This boom is being supported by advances in IoT technologies that enable the collection of big data and machine-learning technologies including deep learning.

Data analysis technologies including machine learning are based on statistical analysis techniques that even today are the target of much research and development activities. Statistical data analysis, however, makes certain assumptions such as normal distributions, so it is known that expected performance cannot be obtained if the data does not follow well-known distributions or if an appropriate distribution is not clear. In response to this problem, Fujitsu Laboratories has been researching topological data analysis (TDA) as a data analysis method that can capture detailed information by focusing on the "shape of data" without using statistical techniques.

Time-series data obtained from sensors or other devices possess the property of "chaos," in which statistical quantities such as mean and variance, frequency, etc., can vary greatly depending on the time of data collection. With this in mind, we have been working to develop technology that can be applied to such time-series data and have developed a technology that combines TDA, deep learning, and anomaly detection technology. We expect this technology to give birth to advanced AI services using sensors.

In this paper, we describe TDA and time-series data analysis technology using TDA and introduce bridge deterioration analysis as an application example.

2. Problem with statistical data analysis

In data analysis including machine learning, conventional statistical analysis techniques make the assumption that data follows some kind of distribution. For example, the mean and deviation of test scores are based on the assumption that the scores follow a normal distribution.

In actual examinations, however, it sometimes happens that the high-score and low-score groups become polarized in a distribution. In this way, there are cases where the distribution the data follows cannot be determined and the information included in the data cannot be described by the probability distribution. In recent years, as the collection of big data became possible, the above-mentioned situation became more frequent, and sufficient performance could not be demonstrated in some cases by conventional statistical analysis methods alone.

To solve this problem, individual analysis methods utilizing data-specific features need to be constructed, or conventional methods need to be expanded to avoid dropping information as much as possible when statistically analyzing a combination of multiple distributions.[1)]

These methods, however, require detailed knowledge of the target data to obtain their unique features. They also require the assumption that the data follows certain distributions locally, which means that the loss of information may be inevitable.

3. Developed technologies in TDA

Research is progressing in the field of TDA as a technology that can obtain a detailed understanding of critical information lost to conventional statistical data analysis by focusing on the "shape of data." This section introduces two key technologies–Mapper and persistent homology–now being researched and developed mainly for TDA.

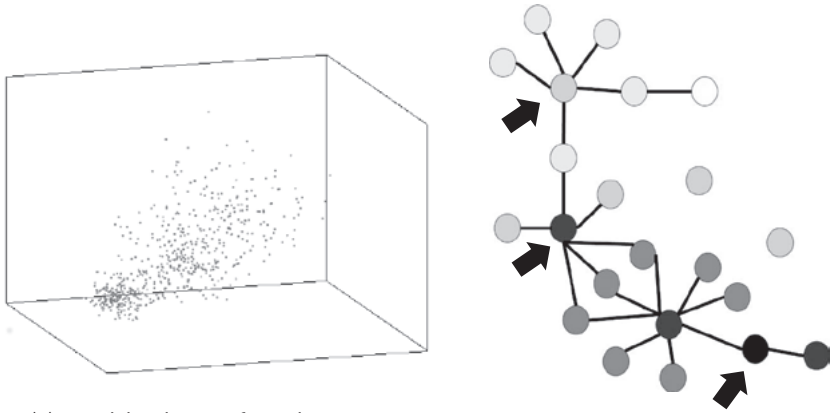
3.1 Mapper

Mapper is a technology that presents the distinguishing features of a set of data as an easy-to-understand graph. To grasp such data features, Mapper groups important parts of that data as nodes and connects nodes having contiguous data by lines (edges), thereby converting that dataset to a graph. Mapper can visualize the distribution of the data by outputting the dataset as a 2D graph.

We can take as an example a distribution of viruses in the case of three sources of outbreak. **Figure 1 (a)** shows the distribution of virus outbreak locations. In the case of a single source of outbreak, the virus would be highly concentrated around the source of the outbreak with its concentration tapering off at a distance from the source. However, in the case of multiple sources of outbreak, the distribution becomes a mixed distribution. Therefore, it would be difficult to tell where the sources are located simply from the distribution map. In this case, Mapper constructs a graph by using data concentration as a basis to group important locations where the concentration is either higher or lower than the periphery into nodes, as shown in **Figure 1 (b)**. In this figure, highly concentrated nodes are shown in dark colors.

It can be seen from Figure 1 (b) that the three nodes indicated by the arrows have a higher concentration than the nodes to which they are connected. This approach enables the user to understand that three sources of outbreak exist without any prior knowledge of that data. This would be difficult to recognize simply by observing the data in Figure 1 (a).

In this way, Mapper is a technology that can easily grasp the shape of data that would otherwise be difficult to understand. Mapper can capture features that are lost in big data and difficult to extract. Because of this capability, it is starting to be applied overseas for a variety of purposes such as anti-money-laundering measures in the field of finance and discovery of associations between disparate diseases in the field of healthcare.[2)]



(a) Mixed distribution of virus locations in the case of three outbreak source

(b) Output by Mapper

Figure 1
Graphing of mixed distribution by Mapper.

3.2 Persistent homology

In contrast to Mapper that can visually grasp the important elements of a certain set of data, persistent homology[3] is a technology that can numerically capture a data shape in detail. It is important in data analysis to understand the arrangement of data, but there are cases in which conventional statistical quantities such as mean and variance cannot convey that feature.

For example, the two sets of data shown in **Figure 2** differ in data arrangement despite having the same mean and variance. In this case, it would not be possible to determine whether a hole exists in the center of the distribution based solely on such quantities. On the other hand, it has been shown that persistent homology will produce a different result for different arrangements of data and will enable information such as a hole in the center to appear as a clear feature. Persistent homology is therefore capable of grasping data features in more detail than conventional statistical quantities.

Persistent homology is outlined in **Figure 3**. This technology considers circles (or spheres in the cases of 3D data) centered about each point of data. When each of these circles expands, the figure can take on a new shape as neighboring circles join up with each other. At this time, the distinguishing characteristics of the data can be understood by observing the change in the number of holes included in the figure. As machine learning commonly uses input data that are composed of equal-length vectors, persistent homology vectorizes the points of data as "Betti sequences." **Figure 4 (a)** is a graph that plots a circle's radius along the horizontal axis and number of connected components (mathematically speaking, zero-dimensional holes) along the vertical axis. **Figure 4 (b)** is a graph that plots the number of ordinary holes (mathematically speaking, one-dimensional holes) along the vertical axis. Circle radius and feature quantities of the figure in each viewpoint from (A) to (D) of Figure 3 correspond to the results in each point of Figure 4 shown by the vertical lines. This graph can be thought of as a means of expressing the arrangement of input data.

Fujitsu Laboratories has developed technology that enables advanced data analysis by combining information quantified by persistent homology with machine learning. It also uses persistent homology in

Figure 2
Example of data having the same mean and variance.

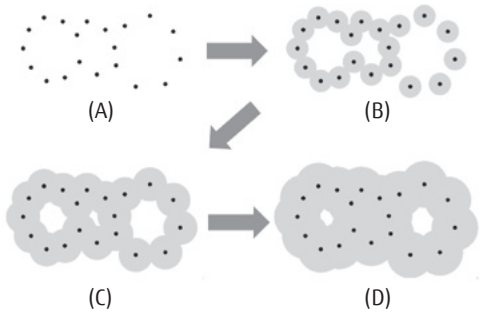
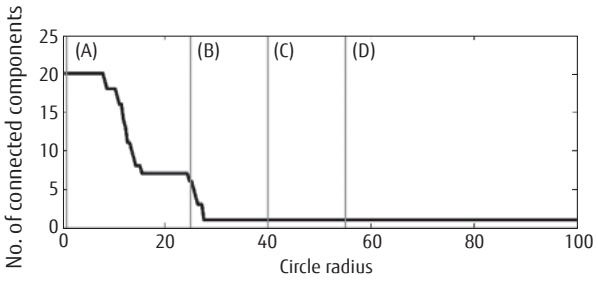
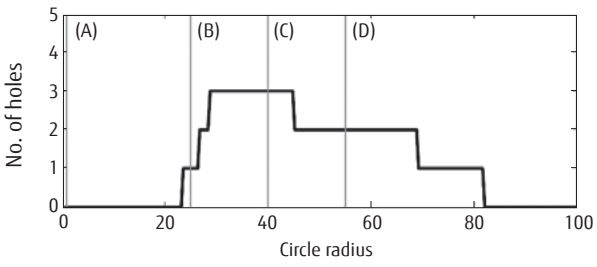


Figure 3
Change in number of expanded circles and holes.



(a) Relationship between circle radius and number of connected components



(b) Relationship between circle radius and number of holes

Figure 4
Visualization of Betti sequences.

relation to technology for converting time-series data. These technologies are described in the next section.

4. Time-series machine learning using TDA

This section introduces two machine-learning technologies using TDA. Known as "supervised learning" and "unsupervised learning," these key technologies provide the two main learning frameworks in the field of machine learning.

4.1 TDA and machine learning targeting time-series data

In machine learning, which is fast becoming the foundation for modern AI, users must shape data beforehand into a form that computers can comprehend. Fujitsu Laboratories is developing AI technologies capable of performing advanced analysis of a wide range of data by shaping data using TDA in combination with technologies like machine learning and deep learning.

To provide some background, recent progress in IoT technologies is making it possible to perform high-performance collection of various types of data from sensors and other devices. In this regard, conventional learning techniques targeting time-series data obtained from sensors have used frequency analysis in addition to statistical quantities such as mean and variance. However, when applying such an approach to intensely fluctuating time-series data, there have been an increasing number of cases in which sufficient performance cannot be achieved. We can consider one cause of this to be as follows: although time-series data possess regularity in their generation, they may exhibit "chaos" in which frequency and statistical quantities like mean and variance are not fixed.

We have developed time-series data analysis technology using TDA as a high-accuracy learning technique for time-series data exhibiting chaos (**Figure 5**). The following summarizes each step.

1) Attractor conversion

Time-series data analysis involves the learning of invariant governing equations as feature quantities even in the case of time-series data having chaos. Governing equations are differential equations that express fluctuation in time-series data. If the generating sources of two sets of time-series data are in the same state, the governing equations are considered to be the same. This approach is therefore considered to be capable of high-performance analysis compared with conventional techniques.

Taking the above into account, we decided to extract information on governing equations using

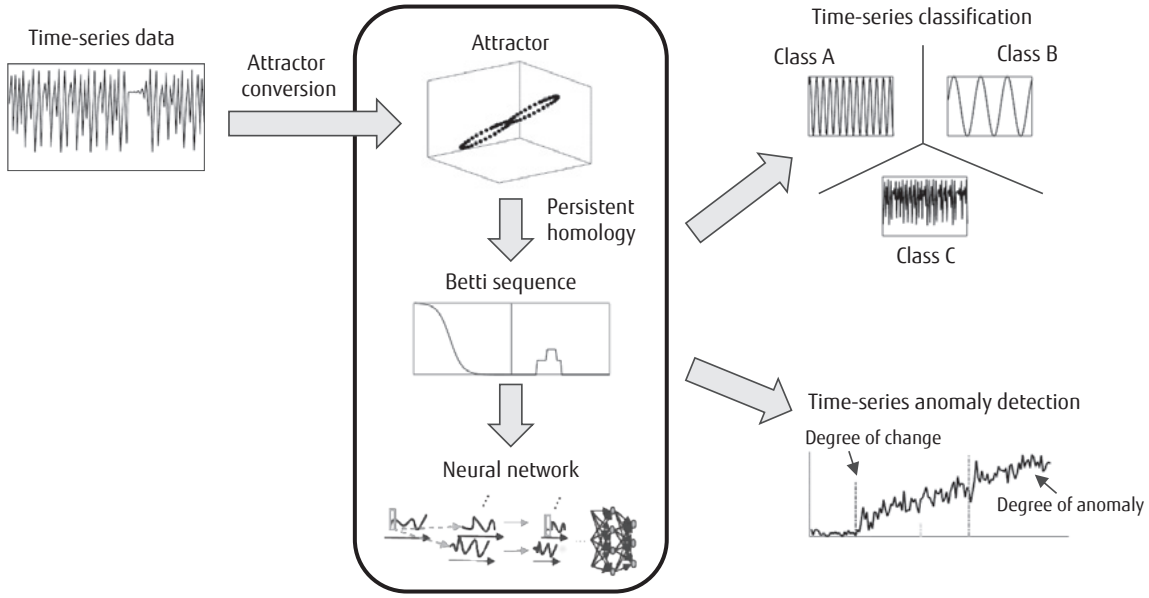


Figure 5
Time-series data analysis technology using TDA.

attractor conversion,[4] which is known in chaos theory as a technique for converting governing equations into a figure. An attractor is a set of solutions to differential equations, so "equation analysis" and "attractor analysis" are synonymous. In addition, considering that time-series data has a finite length in practice, an actual attractor would be a finite set of points.

2) Vectorization by persistent homology

Many techniques in machine learning assume that all input data are composed of equal-length vectors. An attractor configured from time-series data, however, is a set of points, not a vector, so it cannot be used as-is as input for machine learning. For this reason, we adopted persistent homology as a method for vectorizing while preserving the information on point arrangement. Persistent homology makes vectorization easy since it quantifies the information describing an arrangement of points. A vector created in this way is called a "Betti sequence," which can be used as input for machine learning.

3) Supervised learning by one-dimensional convolutional neural network

Given a Betti sequence created from an attractor using persistent homology, a feature component often appears in an adjacent vector element even for time-series data having the same governing equations. This characteristic is analogous to the shifting of an object's position—a feature of that object—in image recognition. A convolutional neural network (CNN) that is highly effective in image recognition can therefore be expected to be effective in the analysis of Betti sequences, so we constructed a CNN especially for this purpose. While image data is two dimensional, a Betti sequence is a one-dimensional vector. There are also cases in which multiple independent data sets are combined, so we constructed an original network that can combine multiple instances of a one-dimensional CNN.[5]

4.2 Evaluation of time-series deep learning

To evaluate time-series deep learning, we performed a comparison experiment with a conventional technique using the following three datasets:

1) Data from gyro sensors attached to both arms, both legs, and the chest of a subject when performing 19 types of actions
2) Brain-wave (electroencephalogram: EGG) data when opening and closing the eyes
3) Muscle-related waveform (electromyography: EMG) data when performing 10 types of actions

Using these three types of data, we compared time-series deep learning with the conventional support vector machine (SVM) algorithm by having each learn feature quantities obtained by conventional statistical processing and feature quantities in chaos theory. We found that time-series deep learning improved the accuracy rate by more than 20% compared with the conventional technique for all of the above datasets, thereby showing the effectiveness of the former as a learning technique for time-series data.

4.3 Time-series anomaly detection technology using TDA

Time-series deep learning is effective if classification is clear and if supervised data affixed with labels in accordance with class are available. However, when performing time-series analysis in actuality, the criteria for classification may not be clear, so the objective often becomes to detect when a change from a normal state to an abnormal state occurs or when signs of such a change occur. We therefore developed anomaly detection technology using TDA.

This technology obtains data from time-series data in fixed periods, converts each set of data to a Betti sequence the same as in time-series deep learning, and compares the results with reference data. The type of reference data used depends on the type of detection being performed. For anomaly detection, data obtained beforehand in a normal state is used, and in the case of change detection, immediately previous data is used as reference data. This algorithm calculates the difference between the target data and reference data as a detection criterion and treats the time at which the difference becomes large as the detection time point. We have prepared multiple methods for calculating the difference between reference and obtained data, including methods based on deep learning. This makes it possible to observe the changes of rules in time-series data generation in the source, which means that this technology is better than conventional techniques at detecting anomalies and changes due to root causes.

4.4 Application of anomaly detection technology using TDA

Fujitsu Laboratories has applied anomaly detection technology using TDA to the analysis of bridge deterioration.

In Japan, many bridges constructed during the period of rapid economic growth are aging. As a result, the business of maintaining and managing these bridges is increasing dramatically, creating problems for society such as rising maintenance costs and an insufficient number of qualified technicians. However, the application of ICT to the business of maintaining and managing social infrastructures such as bridges is expected to help solve these social problems. Against this background, Fujitsu got involved in the Research Association for Infrastructure Monitoring System (RAIMS), where it has taken on the role of accumulating and analyzing monitoring data in relation to the maintenance and management of social infrastructures.

Current bridge inspection work detects and evaluates damage on the basis of up-close visual observation or hammering tests by skilled technicians. However, a visual inspection can only detect deformations appearing on the surface of a structure, that is, it cannot obtain information on the degree of internal damage.

In view of this problem and with the aim of advancing inspection work through the use of ICT, we have undertaken the assessment of bridge damage by attaching sensors on the surface of a bridge deck[note)] and collecting and analyzing vibration data. We have previously used vibration data to assess the degree of internal damage in a bridge deck by a conventional technique such as spectral analysis, but were unable to greatly improve inspection work in this way.

Consequently, to detect and assess internal damage in a bridge deck, Fujitsu decided to apply anomaly detection technology using TDA to monitoring data obtained from acceleration sensors installed on the surface of a bridge deck.[6)] This data was obtained from a moving-wheel load test conducted in FY2015 by RAIMS.

Figure 6 shows the degree of anomaly calculated by anomaly detection technology, the degree of change

note) A basic bridge element that conveys the weight of vehicles traversing the bridge to bridge beams, footing, etc.

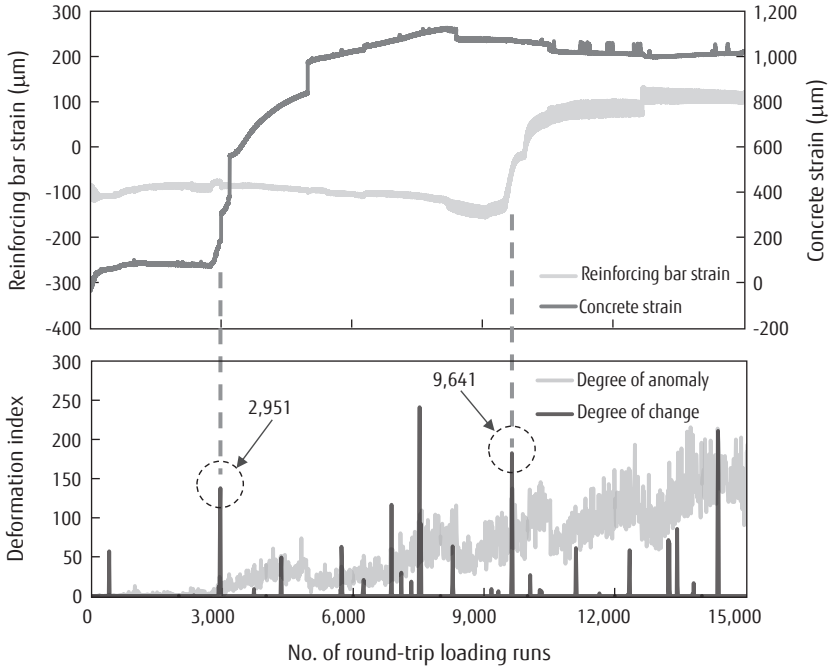


Figure 6
Example of bridge deterioration analysis.

calculated by change detection technology, and the data on internal damage (reinforcing-bar and concrete strain) that drive the deterioration process. The experiment also performed monitoring at some locations using internal strain sensors that cannot be installed in an actual highway bridge. This makes it possible to compare the estimation by the anomaly detection technology and the degree of damage in an actual bridge.

The results of this test show that internal damage occurs at times when the degree of change is large. They also show that the degree of change in anomaly detection can become large at times when no change occurs in the values of internal strain sensors. The reason for this is thought to be that points at which no internal strain sensors were installed were subjected to damaging effects. This result indicates that internal damage in a bridge can be detected at an early stage by applying anomaly detection technology using TDA to data obtained by acceleration sensors installed on the exterior of a bridge deck.

The above application is just one example demonstrating the effectiveness of anomaly detection technology using TDA. Going forward, we plan to apply this technology to actual bridges and to expand its use to detect deterioration in social infrastructures other than bridges. We will also research its application to anomaly detection in mechanical systems and elsewhere.

5. Conclusion

In this paper, we introduced TDA as a new data analysis technique and described time-series deep learning technology and time-series anomaly detection technology developed by Fujitsu Laboratories using TDA.

Going forward, we plan to improve the accuracy of time-series data analysis and develop detailed analysis techniques through joint research with the National Institute for Research in Computer Science and Automation (*Institut National de Recherche en Informatique et en Automatique*: INRIA) in France and to expand TDA beyond time-series analysis to other fields.

References

1) R. Fujimaki et al.: Factorized Asymptotic Bayesian Inference for Mixture Modeling. Proceedings of the 15th International Conference on Artificial Intelligence and Statistics (AISTATS), 2012.
2) Ayasdi: Improving Denial Management Using Machine Intelligence. White paper, 2017.
3) G. Carlsson: Topology and Data. BULLETIN OF THE AMERICAN MATHEMATICAL SOCIETY, Vol. 46, No. 2, pp. 255–308 (2009).
4) A. Basharat et al.: Time series prediction by chaotic modeling of nonlinear dynamical systems. Proceedings of IEEE 12th International Conference on Computer Vision 2009 (ICCV 2009), pp. 1941–1948 (2009).
5) Y. Umeda: Time Series Classification via Topological Data Analysis. Transactions of the Japanese Society for Artificial Intelligence, 32(3), p. D-G72_1–12.
6) J. Kaneko et al.: Study of Monitoring Technology for Fatigue Deterioration in RC Floor Slabs by Moving-wheel Load Test (Part 6)—Evaluation of Fatigue Deterioration by Various Analysis Techniques and Monitoring Data. Japan Society of Civil Engineers, 2017 Annual Meeting (in Japanese).

Yuhei Umeda
Fujitsu Laboratories Ltd.
Dr. Umeda is currently engaged in research and development of machine learning and data analysis.

Junji Kaneko
Fujitsu Laboratories Ltd.
Dr. Kaneko is currently engaged in research and development of mathematical-analysis techniques and their application.

Hideyuki Kikuchi
Fujitsu Laboratories Ltd.
Dr. Kikuchi is currently engaged in research and development of i-Construction (ICT-integrated construction).

©2019 FUJITSU LIMITED