Application of Topological Data Analysis in Complex Network Structure Identification

Zhengqian Lyu *

School of Transportation and Physical Engineering, Shandong Jiaotong University, Ji'nan, China

* Corresponding Author Email: lvzhengqian8866@163.com

Abstract. As a new mathematical tool, Topological Data Analysis (TDA) shows great potential in complex network structure identification. This paper systematically discusses the application method and potential value of TDA in identifying complex network structures. Complex networks, such as social networks, biological networks and traffic networks, are highly nonlinear and dynamic, which challenges the traditional data analysis methods. TDA reveals hidden laws and patterns by digging deep into the internal topological structure of data, which provides a new perspective for network science research. This paper first introduces the basic knowledge of complex network and TDA, and then constructs a comprehensive identification framework, including data preprocessing, topological feature extraction, dimensionality reduction and structure identification. Using persistent homology and other TDA methods, we extract key topological features from the networks and reduce the dimensionality of the feature space using techniques such as Principal Component Analysis (PCA) and t-distributed Stochastic Neighbor Embedding (t-SNE). Finally, classifiers like Support Vector Machines (SVM) are used to identify and classify network structures. Case analysis indicates that the TDA framework can effectively recognize different types of network structures with high classification accuracy.

Keywords: complex network, structure identification, application, topological data analysis.

1. Introduction

As an important tool to describe the structure and behavior of various complex systems in the real world, complex networks have attracted extensive attention. From information dissemination in social networks, gene regulation in biological networks, and traffic management in transportation networks, complex networks are everywhere, and their structural identification plays an irreplaceable role in understanding the internal mechanism of these systems, predicting future behavior and optimizing overall performance [1].

However, the structure of complex networks is often highly nonlinear and dynamic, which makes it difficult for traditional data analysis methods to deal with it effectively. In this context, Topological Data Analysis (TDA), as a new mathematical tool, provides strong support for complex network structure identification with its unique perspective and method [2-3]. TDA can dig deep into the internal topological structure of data and reveal the hidden laws and patterns in complex networks, so it has broad application prospects in the field of network science research.

Although the application of TDA in complex network structure identification has achieved some initial results in recent years, there are still many challenges and problems. For example, how to effectively extract topological features from high-dimensional data, how to improve the accuracy and efficiency of structure identification, and how to deal with the dynamic evolution of networks. The solution of these problems is of great significance to promote the in-depth application of TDA in complex network fields. Therefore, this paper aims to systematically discuss the application method and potential value of TDA in complex network structure identification. By deeply analyzing the topological structure and dynamic characteristics of complex networks and combining with advanced TDA technology, a more efficient and accurate network structure identification method is developed, which provides a new perspective and tool for understanding the nature and behavior of complex systems.

2. Complex network and TDA foundation

Complex network is a huge and highly interconnected structure composed of many nodes and connections between nodes. It is widely used in reality, such as social networks, protein interactive networks and the Internet. The characteristics of this kind of network include large-scale, irregular and non-random self-organizing structure mode, dynamic evolution with time and complex nonlinear behavior of nodes, and its research is helpful to understand and control the complex system behavior in the real world [4].

TDA is a method to study the shape and structure of data, which focuses on revealing the internal laws and patterns of data from a global perspective. Different from traditional data analysis methods, TDA pays more attention to the relative position and connection relationship between data points, rather than the specific value of a single data point [5].

Topological space is an abstract mathematical space, which only cares about the relative position relationship between point sets, without considering the specific distance measurement. In complex network analysis, the network is regarded as a topological space, in which nodes correspond to points in the space and edges represent the connection relationship between points, as shown in Figure 1.

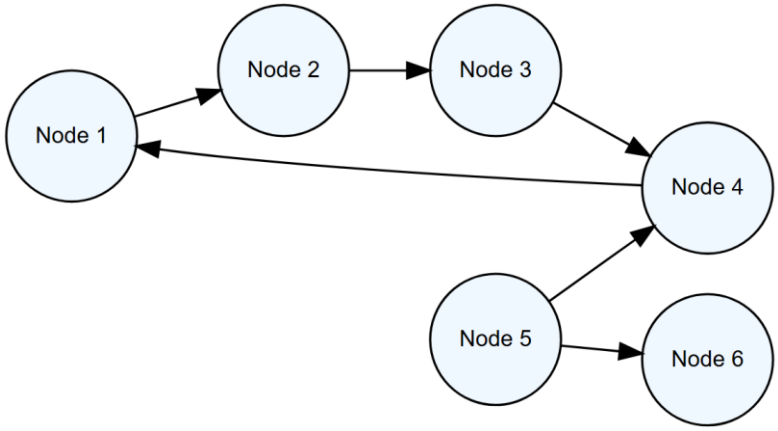

Figure 1. Topological space in complex network analysis

Continuous homology is an important tool in TDA, which is used to quantify the topological characteristics of data at different scales. By constructing the filter complex of data and studying the change of its topological structure with parameters (such as distance threshold), the stable topological features of data are extracted, which are of great significance for data classification, clustering and anomaly detection [6-7]. Mapper algorithm is a tool for visualizing the topological structure of high-dimensional data in TDA. It divides the data into a series of overlapping subsets and applies clustering algorithm to each subset to construct a simplified representation of the data. This method can intuitively understand the overall structure and local details of complex networks.

3. Application method

3.1. Complex network structure identification framework

In order to systematically apply TDA to complex network structure identification, this paper constructs a comprehensive identification framework. The framework mainly includes the following steps, as shown in Figure 2.

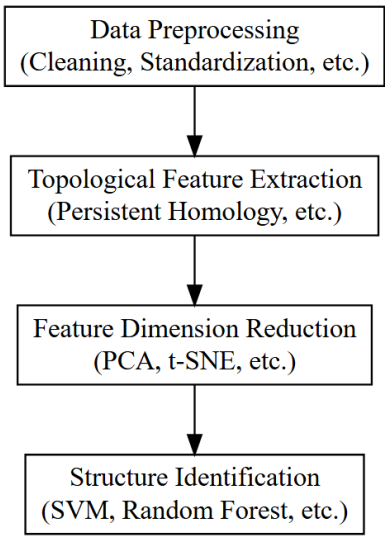

Figure 2. TDA-based framework for complex network structure identification

Firstly, the original complex network data is cleaned, standardized and necessary transformed to ensure the data quality and adapt to the subsequent analysis. Using the methods in TDA, such as continuous homology, the key topological features are extracted from the preprocessed network. These features can capture the global and local structural information of the network.

For a given filtered complex K and scale parameter t, the persistence value of the generator of the k coherent group $H_k(K_t)$ between the appearance time b and the disappearance time d can be defined as:

$$pers(g) = d - b \tag{1}$$

Where g is the generator.

Considering the high dimensionality of the extracted topological features, dimensionality reduction techniques such as Principal Component Analysis (PCA) and t-distributed Stochastic Neighbor Embedding (t-SNE) are employed to reduce the dimensionality of the feature space while retaining the primary information. Support Vector Machines (SVM) are then used to identify and classify network structures utilizing these reduced-dimensional features.

Assuming that the original data matrix is $X \in R^{n\times m}$ (n samples, m features), the goal of PCA is to find the projection matrix $W \in R^{m\times k}$ ($k < m$) to maximize the variance of the projected data $Y = XW$. This can be achieved by solving the eigenvectors corresponding to the first k largest eigenvalues of covariance matrix $C = \frac{1}{n}X^T$.

Given the training data set $\{x_i, y_i\}$ (x_i is the feature vector and $y_i \in \{-1,1\}$ is the category label), the optimization goal of SVM is to find the parameter w, b of hyperplane $w^T x + b$ and the relaxation variable ξ_i, so as to minimize the following objective functions and meet the constraints:

$$\begin{aligned} &\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i \\ &s.t.\ \ y_i\left(w^T x_i + b\right) \geq 1 - \xi_i \\ &\xi_i \geq 0 \end{aligned} \tag{2}$$

Where C is the penalty coefficient.

3.2. Topological feature extraction and dimensionality reduction method

In the aspect of topological feature extraction, the bar code or continuous graph in continuous homology is used to represent the topological features of the network at different scales. Specifically, for a given complex network, the filter complex (for example, rips complex or Čech complex) is constructed, and its persistent homology group is calculated. The generator and death time of these groups constitute a continuous graph, which reflects the stability and changing law of the network structure.

Linear dimension reduction with PCA. PCA projects the data into a low-dimensional space by finding the main direction of change (i.e. principal component) in the data, while maximizing the variance of the projected data. For nonlinear dimensionality reduction, t-SNE is adopted, which is a dimensionality reduction method based on probability distribution and can preserve the local relationship between data points.

3.3. Design and optimization of structural identification algorithm

Choose SVM as the basic classifier. SVM divides different categories of data by finding an optimal hyperplane, while maximizing the interval between the two categories. In order to improve the accuracy of recognition, radial basis function kernel (RBF) is used to map the data to a higher dimensional space, thus enhancing the nonlinear processing ability of the classifier.

An ensemble learning method, such as random forest, is adopted to further improve the stability and accuracy of recognition. Random forest makes the final prediction by constructing multiple decision trees and integrating their outputs. This method can effectively reduce the risk of over-fitting of the model and improve the generalization ability.

4. Case analysis

In this study, the representative data sets including social network, protein interaction network and citation network are selected, and the data are cleaned, standardized and transformed through preprocessing, so as to remove isolated nodes and repeated edges to ensure the accuracy of the data, and adjust nodes and edges to eliminate dimensional differences. Finally, the data are converted into a format suitable for TDA to verify the effectiveness of the TDA framework.

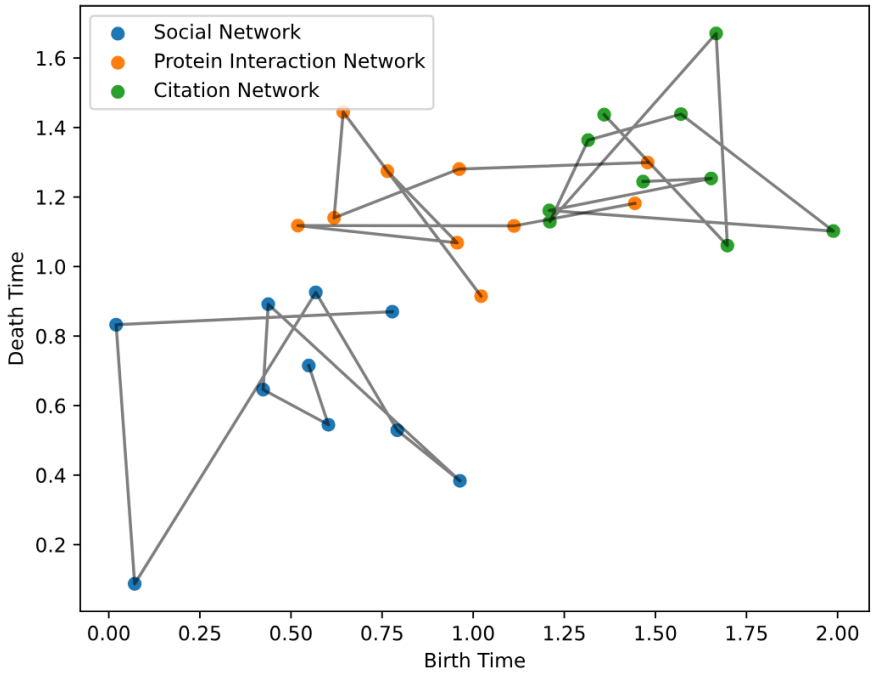

Figure 3. Differences in topological characteristics of different networks

The topological characteristics of different networks show obvious differences (see Figure 3). The characteristics of social networks appear early but mostly disappear quickly, reflecting their dynamic changes. The life cycle of protein interaction network is scattered and long, which indicates that there are more stable interactions and the structure is more complicated. However, the characteristics of citation network appear late and last for a long time, which shows that its citation relationship is lasting and its structure is stable. These differences reveal the uniqueness of different networks in structure and dynamic characteristics, which is helpful to further understand and analyze the characteristics of these networks.

Next, PCA and t-SNE are used to reduce the dimension of the extracted high-dimensional topological features. By comparing the feature space before and after dimensionality reduction, it is found that dimensionality reduction technology can effectively reduce the dimension of features while retaining the main information, thus simplifying the subsequent classification tasks.

Figure 4 shows the data distribution after PCA and t-SNE dimensionality reduction. On the left is the result of PCA dimensionality reduction, and on the right is the result of t-SNE dimensionality reduction. Different network types (social network, protein interaction network and citation network) are distinguished by color. It can be observed from the figure that the data points of different network types are separated to some extent, which shows that PCA effectively retains the main structural information of data and is helpful for the subsequent classification tasks. Compared with PCA, t-SNE keeps the relative distance between data points better, which makes the data points of different network types more clearly separated, which is helpful to evaluate the separability between categories. Dimension reduction technology can effectively reduce the dimension of features while retaining the main information, thus simplifying the subsequent classification tasks.

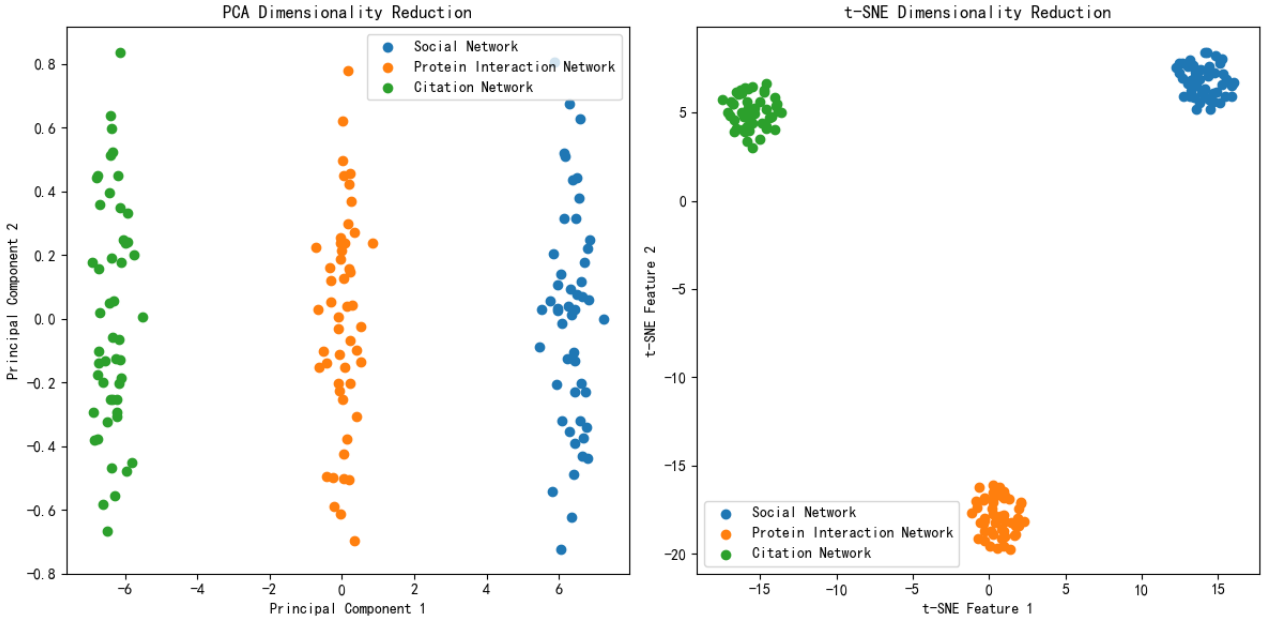

Figure 4. Data distribution after dimensionality reduction by PCA and t-SNE

In the classification stage, SVM is selected as the basic classifier, and RBF is used to enhance its nonlinear processing ability. Different types of network structures are identified and classified by training SVM model. Experimental results show that our method can accurately identify different types of networks structures and has high classification accuracy. However, this method may have some limitations in some specific types of network structure identification. For example, for networks with highly similar topologies, our method may be difficult to distinguish accurately. In order to solve this problem, we plan to further explore more detailed topological feature extraction methods in future research to improve the accuracy and robustness of recognition.

5. Conclusion

The application of TDA in complex network structure identification shows its unique advantages and potential. By digging deep into the internal topological structure of data, TDA can reveal the hidden laws and patterns in complex networks, providing a new perspective and tool for understanding the internal mechanism of these systems, predicting future behavior and optimizing overall performance. In this study, a framework of complex network structure identification based on TDA is constructed, including data preprocessing, topological feature extraction, dimension reduction and structure identification algorithm design and optimization. The example analysis shows that this method can accurately identify different types of network structures and has high classification accuracy. However, for networks with highly similar topological structures, our method may be difficult to distinguish accurately. In order to solve this problem, it is planned to further explore more detailed topological feature extraction methods in future research to improve the accuracy and robustness of recognition.

References

[1] Tong, T., Dong, Q., Yuan, W., & Sun, J. (2024). Identifying vital spreaders in complex networks based on the interpretative structure model and improved kshell. Computing, 106 (5), 1335 - 1358.

[2] Wang, Y., & Zhao, C. (2024). A deep graph convolutional network model of nox emission prediction for coal-fired boiler. The Canadian Journal of Chemical Engineering, 102 (2), 669 - 684.

[3] Isojima, S., Tanioka, K., & Hiwa, H. S. (2023). Preliminary investigation of the association between driving pleasure and brain activity with mapper? based topological data analysis. International journal of intelligent transportation systems research, 21 (3), 424 - 436.

[4] Liu, Y., Wang, J., He, H., Huang, G., & Shi, W. (2021). Identifying important nodes affecting network security in complex networks: International Journal of Distributed Sensor Networks, 17 (2), 1560 - 1571.

[5] Zhukov, M., Hasan, M. S., & Nesterov, P. (2022). Topological data analysis of nanoscale roughness in brass samples. ACS applied materials & interfaces, 14 (1), 2351 - 2359.

[6] Lauric, A., Ludwig, C. G., & Malek, A. M. (2023). Topological data analysis and use of mapper for cerebral aneurysm rupture status discrimination based on 3-dimensional shape analysis. Neurosurgery, 93 (6), 1285 - 1295.

[7] Senge, J. F., Astaraee, A. H., Dotko, P., Bagheri Fard, S., & Bosbach, W. A. (2022). Extending conventional surface roughness iso parameters using topological data analysis for shot peened surfaces. Scientific reports, 12 (1), 5538.