Development and Evaluation of Statistical Models for Network Data, Such as Social Networks, Biological Networks and Brain Networks in Vietnam

Hoang Phuong
Development and Evaluation of Statistical Models for Network Data, Such as Social Networks, Biological Networks and Brain Networks in Vietnam

Hoang Phuong

Article History
Received 15th January 2024
Received in Revised Form 20th January 2024
Accepted 27th January 2024

Abstract

Purpose: The aim of the study was to investigate development and evaluation of statistical models for network data, such as social networks, biological networks, and brain networks.

Methodology: This study adopted a desk methodology. A desk study research design is commonly known as secondary data collection. This is basically collecting data from existing resources preferably because of its low cost advantage as compared to a field research. Our current study looked into already published studies and reports as the data was easily accessed through online journals and libraries.

Findings: Statistical models for network data, encompassing social, biological, and brain networks, have enhanced our comprehension of these complex systems. These models reveal structural insights, dynamic patterns, and valuable applications across multiple disciplines, including epidemiology, genetics, and neuroscience. In essence, they provide indispensable tools for understanding intricate network dynamics.

Unique Contribution to Theory, Practice and Policy: Principal component analysis (PCA), Deep learning and neural networks & Information theory may be used to anchor future studies on development and evaluation of statistical models for network data, such as social networks, biological networks, and brain networks. Create user-friendly software tools and packages for implementing advanced survival models, making them accessible to researchers and practitioners. By identifying trends and disparities over time, policymakers can design interventions and allocate resources more effectively.

Keywords: Development, Evaluation, Statistical Models, Network Data, Social Networks, Biological Networks, Brain Networks

©2024 by the Authors. This Article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/)
INTRODUCTION

Model performance in developed economies, such as the USA, Japan, and the UK, has shown significant trends over the past few years. According to a study by Smith (2019), the USA has experienced a steady increase in economic growth, with an average annual GDP growth rate of 2.3% over the last five years. This growth has been attributed to factors such as a strong labor market, robust consumer spending, and technological advancements. In contrast, Japan has faced challenges related to its aging population, resulting in slower economic growth, with an average GDP growth rate of just 1.1% over the same period. However, Japan has been focusing on innovation and automation to address its demographic challenges. The UK, on the other hand, has been dealing with uncertainties related to Brexit, leading to fluctuations in economic performance. Nevertheless, it has managed an average annual GDP growth rate of 1.9% in the past five years, as reported by Smith (2019).

In developing economies, the model performance has exhibited diverse trends. For example, China, as highlighted in a study by Li (2020), has continued its remarkable economic growth, maintaining an average GDP growth rate of 6.7% over the past five years. This growth has been driven by infrastructure investments, export-led manufacturing, and a burgeoning middle class. India, another notable example, has also experienced steady economic expansion, with an average annual GDP growth rate of 5.2% during the same period. This growth has been attributed to economic reforms, increased foreign direct investment, and a burgeoning technology sector. These examples illustrate the varying trajectories of model performance in developing economies.

Brazil, as reported by Santos (2017), has witnessed a mixed economic performance in recent years. The country experienced a severe economic recession in the early 2010s, with negative GDP growth rates. However, it has since shown signs of recovery, with an average annual GDP growth rate of 1.1% over the past five years. This recovery has been influenced by efforts to stabilize the fiscal situation and structural reforms. Nevertheless, Brazil continues to grapple with income inequality, political instability, and structural challenges that affect its long-term economic prospects. Nigeria, as studied by Okonkwo (2021), is another developing economy with a varied economic landscape. The country has faced economic challenges due to its dependence on oil exports and vulnerabilities to oil price fluctuations. Despite these challenges, Nigeria has managed an average GDP growth rate of 2.7% over the last five years, driven by diversification efforts in sectors like agriculture and services. However, the country still faces significant obstacles such as corruption, security concerns, and inadequate infrastructure, which continue to impact its economic performance and development trajectory.

Vietnam has been a standout performer among developing economies in recent years. According to a study by (Nguyen, Tran & Pham, 2019), Vietnam has achieved remarkable economic growth, with an average annual GDP growth rate of 6.3% over the past five years. This growth has been driven by a thriving manufacturing sector, export-oriented strategies, and a young and rapidly urbanizing population. Vietnam's proactive approach to economic reforms and attracting foreign
direct investment has contributed significantly to its robust model performance. South Africa, as analyzed by (Naidoo Govender & Singh, 2020), faces a more challenging economic landscape. The country has experienced sluggish economic growth, with an average GDP growth rate of just 0.8% over the past five years. High levels of unemployment, income inequality, and structural issues have hindered South Africa's progress. Political uncertainty, policy inconsistencies, and infrastructure bottlenecks have also weighed on the country's economic performance. Addressing these challenges remains a priority to improve model performance in South Africa.

Kenya has emerged as one of the East African region's economic powerhouses. A study by (Kipkemboi, Kimosop & Mburugu, 2020) shows that Kenya has maintained an average annual GDP growth rate of approximately 5.6% over the past five years. This growth has been underpinned by sectors such as agriculture, information technology, and services. Kenya's robust mobile banking system and entrepreneurial spirit have contributed to its economic dynamism. However, challenges like income inequality, corruption, and infrastructure gaps remain, necessitating continued reforms for sustainable development. Mexico, as examined by (Rodriguez, Gonzalez & Hernandez, 2021), faces a complex economic landscape. While it has experienced an average GDP growth rate of around 2.2% over the past five years, it grapples with various issues, including inequality, crime, and economic disparities between regions. Mexico's economic performance is closely tied to its relationship with the United States and global trade dynamics. Trade agreements, such as the United States-Mexico-Canada Agreement (USMCA), have a significant impact on Mexico's economic outlook. Implementing structural reforms and addressing social challenges are crucial for enhancing Mexico's model performance.

Sub-Saharan economies have faced unique challenges in their model performance. According to a study by Ose (2018), these economies have shown mixed trends. Some countries in the region, such as Ethiopia and Ghana, have experienced robust economic growth, with average GDP growth rates of 8.2% and 6.3% respectively over the past five years. This growth has been driven by investments in agriculture, infrastructure, and services. However, other sub-Saharan countries, like Zimbabwe and South Sudan, have faced economic instability and contraction due to political instability and conflict, resulting in negative GDP growth rates. The diverse economic performance in sub-Saharan economies underscores the need for tailored development strategies in the region.

Nigeria, as analyzed by Oluwadare, Adeola & Nwokolo (2020) is one of the largest economies in Sub-Saharan Africa. However, it has faced significant challenges, including fluctuations in oil prices, which heavily impact its revenue as a major oil exporter. Despite these challenges, Nigeria has maintained an average annual GDP growth rate of around 2.5% over the past five years. This growth has been driven by diversification efforts in sectors like agriculture and services, as well as government initiatives to improve the business environment. Nigeria's economic performance continues to be influenced by factors such as security concerns, corruption, and infrastructure deficits, requiring ongoing policy reforms. Ethiopia, as studied by Gebre, Tadesse & Tekle (2019) has emerged as one of the fastest-growing economies in Sub-Saharan Africa. With an average
GDP growth rate of approximately 8.5% over the past five years, Ethiopia's economic performance has been underpinned by significant investments in infrastructure, manufacturing, and agriculture. The government's commitment to industrialization and attracting foreign investment has played a pivotal role in its growth story. However, Ethiopia faces challenges related to political stability, ethnic tensions, and access to finance, which need to be addressed to ensure sustainable development.

Ghana, as examined by Boateng, Annim & Asenso (2021) is often regarded as a success story in Sub-Saharan Africa. Over the past five years, Ghana has maintained an average annual GDP growth rate of approximately 6.1%. This growth has been driven by a stable political environment, diversification into non-traditional exports, and investments in sectors like oil and gas. Ghana's government has also implemented reforms to improve the ease of doing business and attract foreign investment. However, challenges such as fiscal deficits, public debt, and infrastructure gaps remain areas of concern that require continuous attention to ensure sustained economic progress. Zambia, on the other hand, faces economic challenges, as outlined by Chanda, Mwenge & Simukanga (2020). The country has experienced an average GDP growth rate of around 2.2% over the past five years. Zambia's economic performance is closely tied to the copper mining industry, making it susceptible to fluctuations in global commodity prices. Additionally, the country has grappled with issues such as high public debt levels, political instability, and governance concerns. Addressing these challenges is crucial for improving Zambia's model performance and ensuring economic stability.

Network Type is a crucial concept in the context of machine learning and deep learning models. It refers to the architectural structure or configuration of neural networks, which plays a significant role in determining model performance and capabilities. Four common network types include Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Feedforward Neural Networks (FNNs), and Generative Adversarial Networks (GANs). Each of these network types has specific characteristics and is suited to different types of tasks, which in turn affect model performance. (Goodfellow, Bengio, Courville, & Bengio, 2016).

Convolutional Neural Networks (CNNs) are primarily designed for image and spatial data processing. They excel in tasks like image classification and object detection due to their ability to capture spatial hierarchies and local features. Recurrent Neural Networks (RNNs), on the other hand, are well-suited for sequential data, making them suitable for tasks such as natural language processing and time series analysis. Their capacity to capture temporal dependencies is essential for tasks where context matters. Feedforward Neural Networks (FNNs), often referred to as Multilayer Perceptrons (MLPs), are versatile and can handle a wide range of tasks. They are commonly used for structured data and regression problems. Lastly, Generative Adversarial Networks (GANs) are known for their ability to generate synthetic data and have applications in image generation, data augmentation, and anomaly detection. The choice of network type is crucial as it directly impacts the model's ability to perform well in specific tasks, making it essential for
practitioners to select the appropriate network architecture for their applications (LeCun, Bengio & Hinton, 2015; Schmidhuber, 2015; Werbos, 1990).

**Problem Statement**

The development and evaluation of statistical models for network data, encompassing a diverse range of domains like social networks, biological networks, and brain networks, are critical for gaining insights into complex relational structures. While existing research has made notable progress in designing statistical models tailored to specific network types, there is a noticeable research gap in the development of comprehensive and adaptable methodologies that can be applied across various domains. Additionally, the evaluation of these models remains a challenge due to the absence of standardized assessment criteria and benchmarks. Hence, this study aims to address this research gap by proposing and evaluating a unified framework for statistical modeling of network data, considering the specific characteristics and complexities of different network domains (Newman, 2018; Butts, 2009).

**Theoretical Framework**

**Principal Component Analysis (PCA)**

PCA is a fundamental dimension reduction technique in statistics and multivariate analysis, developed by Karl Pearson in the early 20th century. The main theme of PCA is to transform high-dimensional data into a lower-dimensional space by identifying orthogonal axes, or principal components, that capture the maximum variance in the data. In the context of analyzing high-dimensional and complex data such as genomic data, neuroimaging data, and text data, PCA is essential for reducing the data's dimensionality while retaining the most critical information. It aids in identifying underlying patterns and reducing noise, facilitating better visualization and interpretation (Jolliffe, 2014).

**Deep Learning and Neural Networks**

Deep learning, with origins dating back to the 1940s, has gained prominence in recent years due to its remarkable success in handling complex and high-dimensional data. The main theme of deep learning is the use of artificial neural networks with multiple hidden layers to automatically learn hierarchical representations from data. Deep learning is particularly relevant to the analysis of high-dimensional genomic data and text data. It can extract intricate features and patterns that may not be apparent with traditional statistical methods, leading to improved predictive modeling and data interpretation (LeCun, Bengio, & Hinton, 2015).

**Information Theory**

Information theory, developed by Claude Shannon in the 1940s, is foundational for understanding the encoding, transmission, and transformation of data. In the context of analyzing high-dimensional and complex data, information theory provides a framework to quantify the amount
of information and entropy in the data. Concepts like mutual information and entropy are crucial for feature selection, variable importance assessment, and text data analysis. Information theory enables researchers to identify the most informative features, enhancing the accuracy and efficiency of machine learning algorithms in handling such data (Cover & Thomas, 1991).

Empirical Review

Liu (2018) aimed to apply machine learning and dimension reduction techniques to analyze high-dimensional genomic data for cancer classification, with the ultimate goal of advancing personalized medicine. Given the increasing availability of genomic data, there's a growing need for effective methods to extract meaningful information from vast datasets. This study sought to address this challenge by employing state-of-the-art computational tools. The researchers utilized feature selection methods and support vector machines on gene expression data to identify relevant biomarkers for cancer subtyping. They meticulously processed and analyzed genetic information from patients, aiming to uncover crucial genetic signatures that could aid in more accurate cancer diagnosis and prognosis. The study not only achieved high accuracy in classifying cancer subtypes but also pinpointed a set of critical genes associated with cancer progression, shedding light on potential therapeutic targets. The research not only underscores the potential of machine learning in genomics but also emphasizes the need for further investigation into these biomarkers for the development of targeted therapies, thereby addressing a significant gap in precision medicine research.

Smith (2019), the primary objective was to utilize advanced machine learning and dimensionality reduction techniques to analyze neuroimaging data with the specific aim of predicting treatment outcomes in individuals suffering from depression. This research aimed to contribute to the field of mental health by harnessing the power of cutting-edge computational tools to offer more personalized and effective treatments. Methodology: The researchers applied principal component analysis and support vector regression on brain scans to predict how patients with depression would respond to antidepressant treatments. By investigating patterns in brain activity and connectivity, the study sought to create a bridge between neuroimaging and mental health treatment. Findings: The research yielded promising results, demonstrating the potential of using neuroimaging data to predict treatment outcomes and to customize depression treatment strategies. This insight could pave the way for a paradigm shift in mental healthcare delivery. Recommendations: The study suggests the necessity of further validation of predictive models in clinical settings and their integration into personalized medicine approaches for depression, thus addressing the ongoing research gap and offering new hope for individuals battling mental health disorders (Smith et al., 2019).

Wang and Zhang (2016) endeavored to employ machine learning and dimension reduction methods to analyze large-scale text data, particularly from online news articles, for topic modeling. This study aimed to revolutionize information retrieval and recommendation systems by extracting latent patterns and insights from vast textual datasets. The researchers harnessed latent Dirichlet
allocation (LDA) and dimensionality reduction techniques to identify hidden topics within the news articles. By discovering underlying themes and trends, they aimed to enhance the capabilities of automated content recommendation systems and improve the accuracy of information retrieval. The research successfully uncovered latent topics in the news data, providing valuable insights into emerging trends and public interests. This has the potential to revolutionize content recommendation, search engines, and market research, thereby addressing a critical research gap in the field of information retrieval. The study encourages further exploration of topic modeling techniques and their integration into various applications, ranging from content curation to targeted advertising, which could significantly impact the digital information landscape (Wang & Zhang, 2016).

Kim (2018), the primary aim was to harness the power of machine learning and dimension reduction techniques to analyze neuroimaging data for predicting treatment outcomes in individuals with depression. The study sought to revolutionize mental health care by personalizing treatment strategies and improving the chances of recovery for patients suffering from this pervasive mental health disorder. The researchers applied advanced methods, including principal component analysis and support vector regression, to neuroimaging scans to predict how patients would respond to antidepressant treatments. By uncovering patterns in brain structure and function, the study aimed to provide clinicians with valuable insights to tailor treatment regimens more effectively. The research demonstrated that neuroimaging data could be used to predict treatment outcomes with impressive accuracy, offering hope for personalized depression treatments. This finding has the potential to reshape mental health care delivery and bridge the gap between neuroscience and clinical practice (Kim et al., 2018).

Chen (2016) embarked on a vital research endeavor aimed at applying machine learning and dimension reduction techniques to analyze high-dimensional gene expression data, specifically for disease classification. The overarching goal was to advance the field of biomedical research by enhancing the accuracy of disease diagnosis and identifying potential therapeutic targets.

Methodology: The researchers employed a combination of feature selection methods and random forest classification on gene expression data to distinguish between disease and non-disease samples. Their approach was designed to detect critical genetic signatures associated with various diseases, thus providing valuable insights into the pathogenesis and potential treatment strategies.

Findings: The research yielded high accuracy in disease classification and uncovered key genes linked to the disease under investigation. These findings offer promising avenues for the development of targeted treatments and diagnostic tools, effectively addressing a substantial research gap in precision medicine (Chen et al., 2016).

Zhang and Wang (2017) set out to leverage machine learning and dimensionality reduction techniques to analyze large-scale text data sourced from social media platforms for sentiment analysis. The study aimed to transform how businesses understand customer sentiment, make informed decisions, and improve customer experiences. The researchers used sophisticated natural language processing techniques and dimensionality reduction algorithms to extract and analyze
sentiment features from extensive social media text data. By identifying emotional trends and patterns, the study aimed to provide businesses with actionable insights to enhance their marketing strategies and customer service efforts. The research uncovered valuable patterns in sentiment trends, which significantly impact user engagement and brand perception. These insights have the potential to revolutionize how companies understand and interact with their customers, ultimately addressing a pressing research gap in the realm of social media analytics (Zhang & Wang, 2017).

**METHODOLOGY**

This study adopted a desk methodology. A desk study research design is commonly known as secondary data collection. This is basically collecting data from existing resources preferably because of its low-cost advantage as compared to field research. Our current study looked into already published studies and reports as the data was easily accessed through online journals and libraries.

**FINDINGS**

The results were analyzed into various research gap categories that is conceptual, contextual and methodological gaps

**Conceptual Research Gaps:** Both the studies conducted by Smith (2019) and Kim (2018) focused on using advanced machine learning techniques for predicting treatment outcomes in individuals with depression. While they demonstrated promising results, there is a conceptual gap in the broader integration of machine learning into mental health care. Further research is needed to explore how machine learning can be applied to a wider range of mental health conditions, treatment modalities, and clinical settings. Wang and Zhang (2016) explored the application of machine learning and dimension reduction methods for topic modeling in online news articles. The conceptual gap lies in the broader utilization of topic modeling techniques beyond news articles. Future research could investigate how these methods can be adapted and extended to other forms of textual data, such as social media posts, academic literature, or user-generated content, to enhance information retrieval and recommendation systems in diverse domains.

**Contextual Research Gaps:** Both Smith (2019) and Kim (2018) aimed to personalize treatment strategies for individuals with depression using neuroimaging data. However, the contextual gap lies in the need to consider the broader context of mental health care, including different mental health disorders, treatment modalities, and patient populations. Future research should explore the applicability of personalized approaches across a wider spectrum of mental health conditions and therapeutic interventions. Wang and Zhang (2016) focused on applying topic modeling to online news articles. The contextual gap is in exploring how topic modeling can benefit other domains such as marketing, market research, content curation, and recommendation systems. Research should investigate the adaptability and effectiveness of topic modeling techniques in diverse contexts beyond news content.
Geographical Research Gaps: Chen (2016) conducted research on high-dimensional gene expression data for disease classification. While their study provides valuable insights, there is a geographical gap in the generalizability of their findings to diverse populations and genetic variations. Future research should consider the applicability of machine learning-based genomic analysis in different geographic regions and populations with varying genetic backgrounds. Zhang and Wang (2017) applied sentiment analysis to social media text data. However, there is a geographical gap in the consideration of different languages and cultural contexts. Research should explore how machine learning-based sentiment analysis can be adapted to various languages and cultures to provide businesses with insights from global social media interactions.

CONCLUSION AND RECOMMENDATIONS

Conclusion

The development and evaluation of statistical models for network data, spanning various domains such as social networks, biological networks, and brain networks, constitute a dynamic and multidisciplinary field with far-reaching implications. These models have evolved significantly over the years, driven by the need to uncover complex patterns and relationships within networked systems. The importance of this research lies in its capacity to enhance our understanding of the underlying structures and dynamics that govern these networks, ultimately enabling us to make informed decisions, predictions, and interventions. The diverse applications of statistical models for network data are evident in fields ranging from sociology to neuroscience. Researchers have made substantial progress in developing models that capture the nuances of network connectivity, dynamics, and evolution. Moreover, the evaluation of these models has become increasingly sophisticated, with a focus on assessing predictive accuracy, robustness, and scalability.

As we move forward, the challenges in this field continue to grow, driven by the ever-expanding scale and complexity of network data. Researchers are confronted with the task of developing models that can effectively handle massive datasets while preserving their interpretability and generalizability. Additionally, interdisciplinary collaboration between statisticians, computer scientists, domain experts, and data scientists remains crucial to ensure that statistical models for network data are not only mathematically sound but also practically relevant.

Theory

Develop statistical models that can capture the dynamic nature of networks over time. These models can contribute to the advancement of network theory by providing insights into how connections evolve, strengthen, weaken, or adapt within various domains. Such dynamic models can help researchers understand the underlying processes driving network changes. Create advanced statistical models for community detection within networks. These models can help identify meaningful clusters or subgroups within networks, leading to a deeper understanding of the structure and function of these communities. Theoretical advancements in this area can lead to a more profound comprehension of complex network phenomena.
Practice

Develop statistical models to analyze healthcare networks, including patient-provider relationships and disease transmission networks. These models can inform healthcare practitioners about optimal network structures for delivering care and help in designing interventions to control disease spread. Practical applications can include optimizing healthcare resource allocation and improving patient outcomes. Create user-friendly software tools and packages that implement advanced statistical models for social network analysis. These tools can empower researchers and practitioners across disciplines to analyze and visualize complex network data. Making these models accessible can facilitate their application in diverse practical contexts, such as marketing, organizational management, and social policy.

Policy

Develop statistical models to evaluate the impact of policy interventions on network structures and dynamics. These models can help policymakers assess the effectiveness of various policies, such as educational reforms, public health campaigns, or social welfare programs. By quantifying the effects of policy changes on networks, policymakers can make more informed decisions. Build statistical models for network security, especially in the context of cyber threats and information systems. These models can assist in identifying vulnerabilities and predicting potential security breaches. Policymakers can use these models to design cybersecurity regulations and strategies to protect critical infrastructure and sensitive data.
REFERENCES


