Selected Ph.D. Thesis Abstracts

This Ph.D thesis abstracts section presents theses defended in 2023 and 2024. These submissions cover a range of research topics and themes under intelligent informatics, such as methods and techniques for counterfactual explainability in graphs, addressing the challenges in using feature engineering to classify cancer, approaches to preserving privacy in graph machine learning, and techniques for self-consistent explanation in mental status computation.

Counterfactual explainability in graphs: foundations, generative methods, and ensembles techniques

Mario Alfonso Prado Romero
marioalfonso.prado@gssi.it
Gran Sasso Science Institute, Italy

Graph Neural Networks (GNNs) have demonstrated exceptional performance across various domains, including social network analysis, recommender systems, and biochemistry. These models excel at learning from graph-structured data, where both the individual attributes of nodes and their relationships (edges) are crucial for predictions. However, GNNs operate as black-box models, which lack interpretability, a critical limitation in fields like healthcare and finance where understanding decision-making processes is essential for building trust and ensuring fairness.

Graph Counterfactual Explanations (GCEs) offer a promising approach to explaining GNN predictions by revealing how slight changes to the graph's structure or node features could lead to different outcomes. Unlike traditional feature-based explanations, counterfactuals provide actionable insights, suggesting concrete steps users can take to achieve a desired outcome. In fields like healthcare, GCEs could explain how a patient might lower their risk of disease, while in finance, they might suggest how to improve creditworthiness for a loan.

Despite their potential, generating effective GCEs presents several challenges: i) the absence of a general definition and taxonomy of GCE approaches, ii) the lack of standardized evaluation methods - as current approaches use diverse datasets, metrics, and oracles - complicates performance comparisons, iii) ensuring that counterfactuals are both plausible (i.e., realistic within the data distribution) and actionable (i.e., providing practical steps), and iv) the intrinsic nature of graph data demands explanations that incorporate both nodes' and relationships' attributes, which are typically intertwined and domain-specific.

This thesis tackles the challenges mentioned above by advancing the SoA through several key innovations. First, we tackled i) and ii) by rationalizing the field of study in a thorough Survey, where we shaped a general definition of GCE and an exhaustive taxonomy of the existing approaches. Beyond that, we provided a qualitative comparison, complemented by an empirical comparison, of existing graph counterfactual explanation methods. Furthermore, we analyzed the most widely used datasets and evaluation metrics in the GCE's SoA, assessing their strengths and limitations. This foundational work, by presenting a unifying perspective that facilitates interpreting and comparing various GCE methods, provided the clarity much needed by the nascent and fragmented field of study of GCE. Second, we set a milestone for ii) by introducing GRETEL, a versatile and extensible framework designed for developing and fairly evaluating GCE methods. This framework offers a comprehensive set of tools, including datasets, oracles, explainers, and evaluation metrics, to standardize and streamline the evaluation process. Third, we contributed to solving iii) by proposing a novel Generative GCE method that produces plausible counterfactuals aligned with the underlying data distribution, ensuring that the explanations are realistic within the context of the problem domain. Lastly, we faced challenges iv) by exploiting ensemble methods that, through selection and aggregation strategies, combine multiple GCE techniques to provide more robust counterfactuals, thus enhancing performance in the diverse datasets from different domains.

Overall, by enhancing the transparency of model behavior, the contributions presented in this thesis will enable users to make well-informed decisions based on AI-driven predictions in critical domains, including healthcare, finance, and social network analysis.

Feature Engineering for Cancer Data Modeling

Markian Jaworsky
markian.jaworsky@gmail.com
University of Southern Queensland, Australia

A range of risk factors increases the likelihood of developing chronic illness, awareness, and understanding of these possible causes, give patients the best chance of survival by making informed life choices. Finding patterns amongst risk factors and chronic illness can suggest similar causes and provide guidance information to improve healthy lifestyles, and where outliers appear, gives clues for possible treatments. Prior studies have typically isolated data challenges of single disease datasets, however, to establish a truly healthy lifestyle the predictive feature power of many diseases is more useful. We discuss the 4 most common data challenges in health surveys and propose a novel approach to the selection of features to optimize a multi-label classifier of diabetes and 30 types of cancer, to establish a total healthy lifestyle. A novel knowledge graph is constructed from the text of health survey questions and used to determine the weight of feature relationships based on World Health Organization

(WHO) ICD codes, to prioritize selection. The results of our study demonstrate that our knowledge of graph-based feature selection, when applied to several machine learning and deep learning multi-label classifiers, improves precision, recall, and F1 score.

Six known cancer types positively correlate to diabetes diagnosis. Diabetes is known to increase the risks of pancreatic, liver, colorectal, breast, endometrial, and bladder cancer diagnosis and outcomes. The study highlights that the characteristics of diabetes, including high blood sugar and insulin levels, and inflammation are also known risk factors for cancer cells to proliferate, grow, and metastasize. This understanding can help patients determine their cancer risk as being a combination of exposure to risk factors that specifically trigger DNA damage, and secondly create inflammation in the human body that promotes cancer cells to thrive. Alternatively, diabetes decreases the risk of prostate cancer, further understanding of this negative correlation can potentially assist with prostate cancer treatment or prevention.

There exists an abundance of research papers that aim to solve data challenge scenarios in the Feature Engineering for Cancer Data Modeling analysis of health survey data, but rarely presented is a framework designed to address the many combined challenges. An exception to this trend is the multiple contributions of other researchers, who propose a novel feature selection approach to handling high dimensional imbalanced class data and designing a classifier to address the issue of data nonlinearity. A shortcoming of this paper is the preparation of a dataset that most likely will contain samples with features consisting of missing values. The study concluded that outliers were a challenge, yet this additional scenario was to be handled.

This research proposal seeks to take a holistic approach to the classification of cancer, addressing all data challenges presented in the world's largest annual health survey, and attempting to create a multi-label classifier of up to 30 individual cancer subtypes, multiple cancer occurrences, and their association with diabetes.

Privacy Preserving Graph Machine Learning

Olatunji Iyiola Emmanuel
iyiola_olatunji@yahoo.com
Leibniz Universität Hannover, Germany

Most real-world data can be represented as graphs, capturing intricate relationships and dependencies among entities. This unique characteristics of graphs makes them applicable in various domains. A special family of machine learning models called graph neural networks (GNNs) are specially designed to handle graph data. In recent years, the widespread adoption of GNNs has revolutionized various analytical tasks involving graph data, such as node classification and link prediction. However, concerns regarding the privacy vulnerabilities of these models have emerged, particularly in sensitive domains like healthcare, finance and recommender systems. This thesis explores the privacy implications of GNNs through a multi-faceted analysis encompassing several attacks, defense strategies and privacy-preserving frameworks.

The first investigation focuses on the susceptibility of GNNs to membership inference attacks. We propose several attacks and defenses to effectively mitigate these attacks while minimizing the impact on model performance. Our findings reveal that structural information rather than overfitting is the primary contributor to information leakage.

Subsequently, we propose a novel privacy-preserving framework, \pkg, leveraging knowledge distillation and two noise mechanisms, random subsampling, and noisy labeling to privately release GNN models while providing rigorous privacy guarantees. The theoretical analysis within the Rényi differential privacy framework is accompanied by empirical validation against baseline methods. We also show that our privately released GNN model is robust to membership inference attacks.

Furthermore, since model explanations have become a desirable outcome of modern machine learning models, we explore the privacy risks involved in releasing model explanations from GNNs. Specifically, we study the interplay between privacy and interpretability in GNNs through graph reconstruction attacks. We demonstrate how model explanations can facilitate the reconstruction of sensitive graph structures. Various attack strategies are evaluated based on auxiliary information available to adversaries, with a proposed defense employing randomized response mechanisms to mitigate privacy leakage.

Lastly, we develop attacks to systematically study the information leakage from latent representation in graph and tabular input data domains. We reveal the susceptibility of latent space representation learning to privacy attacks that reconstruct original input with high accuracy. Furthermore, we utilize these attacks as privacy auditors to evaluate the privacy guarantees of differentially private models on both graph and tabular data, providing valuable insights into the privacy risks associated with releasing latent space representations.

By comprehensively addressing these privacy challenges, this thesis contributes to a deeper understanding of the privacy implications of GNNs and provides practical insights into enhancing their privacy-preserving capabilities in real-world applications.

Link to the official soft-copy version:
https://www.repo.uni-hannover.de/handle/123456789/18093

Website: https://iyempissy.github.io

Towards Self-Consistent Explanation for Mental Status Computation

Xiaohua Wu
xhwu@whut.edu.cn
Wuhan University of Technology, China

Mental status is an important psychological characteristic, representing an individual's experience of a particular moment, situation, or condition. Computing an individual's mental status can play a significant role in the fields of education, security, human resource management, psychiatric diagnosis, and social decision-making. The current dominant

approaches are regression-based and conventional machine learning-based (ML) methods, often applied to data collected from online question- answering communities. While these methods have achieved acceptable prediction performance with strong interpretability, their modeling ability is limited, making it challenging to extract key factors of mental status in the era of large-scale social media data. Previous work has demonstrated the effectiveness of deep learning in mental status computation. However, deep learning-based approaches are commonly considered a "black box" due to their lack of interpretability, which restricts their use in critical areas such as psychiatric diagnosis, security, and socially assisted decision-making.

In this thesis, we propose a self-consistent explanation for mental status computation and focus on the following three problems. The first topic explores how to quantitatively explain the factors involved in mental status computation. Current explainable methods often overlook factor interactions and struggle to quantify each factor's contribution. In response, we employ Shapley value, a post-hoc explanation method based on cooperative game theory with strong theoretical support for quantifying factor contributions. Additionally, a post-hoc explanation method combining multi-factor interaction is designed.

The second topic addresses how to achieve self-consistent explanation results across different models when model-agnostic explanation methods are used to calculate factor explanations. This self-consistency is of benefit to trustworthy decision-making. We propose a multi-model joint factor contribution computation method to obtain consistent interpretation results, which in turn guides model training, leading to significant improvements in mental status computation.

Finally, we identify potential problems with the Shapley value, particularly in the construction of samples that are unlikely to appear simultaneously in real-world applications due to random bias. To address this, we propose an optimized Shapley value that removes contained pseudo-factor samples based on co-occurrence and mutual exclusion constraints. This enhancement improves both the computational efficiency of the explanation method and the consistency of the explanation results with real-world scenarios.