Integrative Network Analysis: Unveiling Symptom-Disease Interactions and Enhancing Predictive Models

Andreoli C. • Ligari D. • Alberti A. • Scardovi M. [1]

[1] *Department of Computer Engineering, Data ScienSickness predictionce, University of Pavia, Italy*
Course of Financial Data Science

Github repository: https://github.com/AndreaAlberti07/enhancing-disease-prediction

Abstract

We will write it once we have the results.

Keywords— *Graph theory • Features Engineering • Community detection • Null models • Random forest • MLP*

Contents

1 Introduction

In the dynamic landscape of healthcare, understanding the intricate interplay between symptoms and diseases is paramount for effective diagnosis and prediction. This report embarks on a comprehensive journey through the realms of network analysis, leveraging both theoretical foundations and empirical data to unravel the complexities of symptom-disease interactions. Our dual-fold objective is to provide a nuanced descriptive analysis of these interactions while identifying key features to bolster predictive models.

The foundation of this endeavor lies in an extensive review of existing literature, drawing insights from seminal works on network theory and disease prediction. By establishing a baseline through prior research, we pave the way for a deeper understanding of the subject matter and ensure the relevance of our findings in the broader context of scientific inquiry.

Guided by insights gleaned from the literature, our exploration extends to the realm of data, where we meticulously curate and analyze datasets of varying sizes. Through a systematic process of exploratory data analysis and cleaning, we prepare the groundwork for constructing meaningful networks that encapsulate the relationships between symptoms and diseases.

The heart of our analysis lies in the creation of intricate network structures, employing bipartite models and non-weighted links to distill meaningful patterns. We delve into a spectrum of network metrics, from fundamental measures like degree distribution and clustering coefficients to more nuanced assessments of node importance and betweenness centrality. Statistical significance is rigorously assessed through the lens of a null model, ensuring that our observations transcend mere chance.

Community detection algorithms further dissect the network, revealing hidden structures and relationships between diseases. This not only enriches our understanding but also lays the groundwork for subsequent analyses. As we traverse the terrain of network analysis, we introduce novel metrics inspired by the Hidalgo-Hausmann framework, stratifying symptoms and diseases based on their predictive importance. These metrics, coupled with traditional measures like betweenness centrality, contribute to the definition of features that fuel our predictive models.

With a robust foundation established, we transition to the realm of predictive modeling, where our feature-rich approach promises to enhance the performance of established models. Logistic regression, random forest, and multi-layer perceptron models are trained, tested, and validated, with a keen eye on feature importance and model improvement strategies.

This report unfolds as a holistic exploration, weaving together theoretical frameworks, empirical analyses, and predictive modeling into a cohesive narrative. As we traverse the intricate web of symptom-disease interactions, our aim is not only to elucidate the underlying dynamics but also to pave the way for more accurate and insightful predictive models in the realm of healthcare.

1.1 Network Creation (Not Weighted - Bipartite)

1.2 L1 and L2 measures

1.3 Clustering

To compute the average network clustering coefficient, according to Watts and Strogatz 1998, Watts and Strogatz, it is possible to use the following formula:

$$C = \frac{1}{n}\sum_{i=1}^{n} C_i = \frac{1}{n}\sum_{i=1}^{n} \frac{2e_i}{k_i(k_i-1)} \tag{1}$$

where n is the number of nodes, C_i is the clustering coefficient of node i, e_i is the number of edges between the neighbors of node i and k_i is the degree of node i. Specifically, we used the version of clustering coefficient for bipartite graphs, redefined by Latapy, Magnien, and Vecchio 2008, Latapy et al. and implemented in the NetworkX function `nx.bipartite.average_clustering`.

1.4 Betweenness Centrality

1.5 Communities with Co-occurrence Matrix

1.6 Degree Distribution and Power Law

1.7 Most Important Symptoms/Diseases (4 Classes)

1.8 Communities

1.9 L2 Has No Sense, It's Right That Z-Score Low Value

1.10 Meaning of Z-Score

1.11 Betweenness Centrality

1.12 Clustering Coefficient

1. Average Clustering Coefficient for the Entire Bipartite Graph (0.114): - Indicates a moderate level of local clustering in the entire network, capturing the tendency of symptoms and diseases to form clusters.

2. Average Clustering Coefficient of Diseases (0.132): - Diseases exhibit a higher clustering coefficient compared to the overall graph. - Diseases are more interconnected with common symptoms, forming localized clusters in the network.

3. Average Clustering Coefficient of Symptoms (0.071): - Symptoms have a lower clustering coefficient compared to the overall graph. - Symptoms are less likely to form tightly connected clusters among themselves.

Analysis: - Diseases show a stronger tendency to share common symptoms and form clusters, contributing to the higher clustering coefficient observed for diseases. - Symptoms, on the other hand, exhibit a more dispersed pattern, indicating that common symptoms may not necessarily co-occur with each other at a high frequency.

Conclusion: - The network's clustering patterns suggest a structured organization, with diseases playing a central role in forming clusters based on shared symptoms. - The lower clustering coefficient for symptoms implies greater heterogeneity among symptoms, emphasizing the need for careful consideration of diverse symptom profiles in disease prediction. - These findings align with the bipartite nature of the network, highlighting the meaningful connections between diseases and symptoms that contribute to the overall clustering patterns.

References

Devi, G. U. 2015. "Detection of DDoS Attack using Optimized Hop Count Filtering Technique" [in en]. *Indian Journal of Science and Technology* 8, no. 1 (January): 1–6. ISSN: 09746846, 09745645. https://doi.org/10.17485/ijst/2015/v8i26/83981.

Fang, L., H. Wu, K. Qian, W. Wang, and L. Han. 2021. "A Comprehensive Analysis of DDoS attacks based on DNS." *Journal of Physics: Conference Series* 2024:012027. https://doi.org/10.1088/1742-6596/2024/1/012027.

Latapy, M., C. Magnien, and N. D. Vecchio. 2008. "Basic notions for the analysis of large two-mode networks." *Social Networks* 30, no. 1 (January): 31–48. ISSN: 0378-8733. https://doi.org/10.1016/j.socnet.2007.04.006.

Taylor, R. 2021. *Four major DNS attack types and how to mitigate them* [in en-US], August. Accessed May 10, 2023. https://bluecatnetworks.com/blog/four-major-dns-attack-types-and-how-to-mitigate-them/.

Watts, D. J., and S. H. Strogatz. 1998. "Collective dynamics of 'small-world' networks" [in en]. *Nature* 393, no. 66846684 (June): 440–442. ISSN: 1476-4687. https://doi.org/10.1038/30918.