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 Betweenness Centrality

The betweenness centrality of a node v, according to Brandes [2], is defined as the sum of the fraction of all-pairs shortest paths that pass through v:

$$c_B(v) = \sum_{s,t \in V} \frac{\sigma(s,t|v)}{\sigma(s,t)} \qquad (1)$$

where:

- V: The set of nodes.
- $\sigma(s,t)$: The number of shortest paths from node s to node t.
- $\sigma(s,t|v)$: The number of those shortest paths from node s to node t that pass through some node v other than s and t.
- If $s = t$, then $\sigma(s,t) = 1$.
- If $v \in \{s,t\}$, then $\sigma(s,t|v) = 0$.

To compute the betweenness centrality we used the NetworkX function `nx.bipartite.betweenness_centrality` which implements the algorithm proposed by Brandes [1] and uses a proper normalization for bipartite graphs.

1.4 Communities Detection

Prior to apply any community detection algorithm, we need to perform two steps:

- **Graph Projections**: We need to project the bipartite graph into two graphs, one for each set of nodes. In our case the two sets are represented by symptoms and

diseases. At this scope is available the NetworkX function `nx.bipartite.projected_graph` which returns the projection of the bipartite graph onto the specified nodes.
- **Compute Similarity**: We need to compute the similarity between nodes. For our purposes, it is possible to create a co-occurrence matrix, for each set of nodes. Taking as example the co-occurrence matrix of symptoms, each entry s_{ij} represents the number of times the symptom i and the symptom j co-occur in the same disease.

Once we have the two graphs, whose links are weighted by the similarity between nodes, we can apply the community detection algorithm. We used the Clauset-Newman-Moore greedy modularity maximization algorithm [3], implemented in the NetworkX function `nx.algorithms.community.greedy_modularity_communities`. This algorithm aims at finding the partition of the graph that maximizes the modularity, which is defined by Newman [6] as:

$$Q = \frac{1}{2m} \sum_{ij} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j) \tag{2}$$

where:

- Q: Modularity of the network.
- A_{ij}: Element of the adjacency matrix representing the connection between nodes i and j.
- k_i and k_j: Degrees of nodes i and j, respectively.
- m: Total number of edges in the network.
- $\delta(c_i, c_j)$: Kronecker delta function, which is 1 if c_i is equal to c_j (i.e., nodes i and j belong to the same community) and 0 otherwise.
- The sum is taken over all pairs of nodes i and j.

1.5 Degree Distribution and Power Law

1.6 Most Important Symptoms/Diseases (4 Classes)

1.7 Communities

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

1.9 Meaning of Z-Score

1.10 Betweenness Centrality

References

[1] Ulrik Brandes. “A Faster Algorithm for Betweenness Centrality”. In: *The Journal of Mathematical Sociology* 25 (Mar. 2004). DOI: 10.1080/0022250X.2001.9990249.

[2] Ulrik Brandes. “On variants of shortest-path betweenness centrality and their generic computation”. In: *Social Networks* 30.2 (May 2008), pp. 136–145. ISSN: 0378-8733. DOI: 10.1016/j.socnet.2007.11.001.

[3] Aaron Clauset, M. E. J. Newman, and Cristopher Moore. “Finding community structure in very large networks”. In: *Physical Review E* 70.6 (Dec. 2004). arXiv:cond-mat/0408187, p. 066111. ISSN: 1539-3755, 1550-2376. DOI: 10.1103/PhysRevE.70.066111.

[4] G. Usha Devi. “Detection of DDoS Attack using Optimized Hop Count Filtering Technique”. en. In: *Indian Journal of Science and Technology* 8.1 (Jan. 2015), pp. 1–6. ISSN: 09746846, 09745645. DOI: 10.17485/ijst/2015/v8i26/83981.

[5] Lei Fang et al. “A Comprehensive Analysis of DDoS attacks based on DNS”. In: *Journal of Physics: Conference Series* 2024 (2021), p. 012027. DOI: 10.1088/1742-6596/2024/1/012027.

[6] M. E. J. Newman. “Modularity and community structure in networks”. In: *Proceedings of the National Academy of Sciences of the United States of America* 103.23 (June 2006), pp. 8577–8582. ISSN: 0027-8424. DOI: 10.1073/pnas.0601602103.

[7] Rebekah Taylor. *Four major DNS attack types and how to mitigate them*. en-US. Aug. 2021. URL: https://bluecatnetworks.com/blog/four-major-dns-attack-types-and-how-to-mitigate-them/ (visited on 05/10/2023).