Exploring Gene Causality

Name – Abhishek S Waghchaure

Candidate ID - NaukriLtr092024

1. Introduction

The goal of this analysis was to investigate potential causality relationships between genes and phenotypes using embeddings generated for both entities. The exercise focused on leveraging techniques such as dimensionality reduction, clustering, and vector analysis to identify and analyze relationships that may indicate a gene's causal role in a phenotype. This report presents the methodology, results, and key insights obtained through the analysis.

2. Data Overview

We worked with four primary datasets:

1. **Phenotype-Gene Pairs (**opentargets_step2.for_llm.tsv**)**: This dataset contains the relationships between phenotypes and associated genes, providing the basis for understanding which genes are linked to which phenotypes.
2. **Causal Gene Labels (**opentargets_step2.labels**)**: This dataset identifies the causal gene for each phenotype from the associated gene list, enabling us to differentiate causal from non-causal genes.
3. **Gene Embeddings (**gene_embeddings.csv**)**: Contains high-dimensional embeddings (3072 dimensions) for each gene, representing the gene's semantic and functional properties.
4. **Phenotype Embeddings (**phenotype_embeddings.csv**)**: Similar to gene embeddings, this file contains 3072-dimensional vectors representing each phenotype.

3. Methodology

3.1 Data Preparation

The first step was to load and clean the provided datasets. The phenotype-gene pairs were mapped, and the causal gene for each phenotype was identified using the opentargets_step2.labels dataset. This labeling process allowed us to differentiate between causal and non-causal genes for each phenotype.

To make the dataset unique for the exercise, we sampled 500 phenotype-gene pairs using a hash of my name (“Abhishek Waghchaure”), ensuring reproducibility and uniqueness.

3.2 Dimensionality Reduction (PCA)

Both gene and phenotype embeddings were high-dimensional (3072 dimensions). To visualize and explore the relationships between these entities, we applied **Principal Component Analysis (PCA)** to reduce the dimensionality to 2 components.

- **PCA on Gene Embeddings**: Helped in visualizing gene clusters.
- **PCA on Phenotype Embeddings**: Allowed for a better understanding of how phenotypes group together.

3.3 Clustering Analysis

We applied three different clustering techniques to group the gene and phenotype embeddings:

- **K-Means Clustering**:
- Used the **Elbow Method** to identify the optimal number of clusters.
- Evaluated clustering performance using **Silhouette Score** and **Davies-Bouldin Index**.
- **Hierarchical Clustering**:
- Generated a **dendrogram** to explore how genes and phenotypes merge into clusters hierarchically.

- **DBSCAN**:
- Tried density-based clustering to identify regions of high density in the gene and phenotype space.
- After parameter tuning, it was found that DBSCAN struggled to form meaningful clusters due to the distribution of the embeddings.

3.4 Vector Analysis

The final step in the methodology was to calculate the **difference vectors** between the embeddings of genes and phenotypes. This vector analysis was intended to quantify the proximity of causal and non-causal gene-phenotype pairs in the embedding space.

- **Purpose**: To determine if the difference in embeddings between a gene and a phenotype could be a proxy for causality.
- **Future Use**: These difference vectors can be used as features for supervised learning models to classify gene-phenotype pairs as causal or non-causal.

4. Results

4.1 PCA Visualization

- The PCA-reduced embeddings were plotted to visualize the relationships between genes and phenotypes.
- **Gene Embeddings**: The scatter plot showed moderate clustering of genes.
- **Phenotype Embeddings**: Similarly, phenotypes showed some clustering tendencies, although the separation wasn't clear-cut.

4.2 Clustering Performance

- **K-Means Clustering**: Using the Elbow Method, we identified 3 clusters as optimal. The clusters formed were moderately well-separated based on **Silhouette Scores (~0.39)** and **Davies-Bouldin Index (~0.84)**, indicating decent but not perfect clustering.
- **Hierarchical Clustering**: The dendrogram provided a clear hierarchical structure of the data, but the clusters formed at different levels were less distinct than with K-Means.

- **DBSCAN**: Despite tuning, DBSCAN failed to form meaningful clusters due to the uniformity and spread of the embeddings. Most points were classified as noise, suggesting DBSCAN might not be suitable for this type of data.

4.3 Vector Analysis

- **Difference vectors** between gene and phenotype embeddings were calculated, with the assumption that causal relationships should result in smaller differences (i.e., closer embeddings).
- These difference vectors can be useful for future classification tasks, where a model could learn to differentiate causal from non-causal gene-phenotype pairs based on these vectors.

5. Challenges & Limitations

- **DBSCAN Clustering**: The embeddings used for the analysis did not naturally form dense clusters, which caused DBSCAN to label most points as noise, even after parameter tuning.
- **Clustering Quality**: The **Silhouette Scores** and **Davies-Bouldin Index** suggested that while clusters were formed, they were not tightly separated. This could indicate that the embeddings themselves may not fully capture the causal relationships between genes and phenotypes.
- **Interpretation of Clusters**: Further biological interpretation of the clusters would be needed to determine if they have real-world significance in terms of gene-phenotype relationships.

6. Conclusion

The analysis successfully explored the relationships between genes and phenotypes using clustering and vector analysis. The key findings include:

- **K-Means** was the most effective clustering method for the given embeddings, yielding moderate clustering quality.
- **Vector analysis** provided a potential framework for future supervised learning tasks by quantifying the difference between gene and phenotype embeddings.

- **Dimensionality reduction (PCA)** helped visualize the high-dimensional space and provided a clearer understanding of how genes and phenotypes relate.

While the results are promising, further exploration and refinement of the embeddings, as well as more advanced models, may be required to uncover stronger causal relationships.

7. Future Work

- **Supervised Learning**: Use the calculated difference vectors between genes and phenotypes as features for a classifier to predict causality. This would be a natural extension of the vector analysis.
- **Improved Embeddings**: Consider using alternative embedding techniques or fine-tuning existing embeddings to improve clustering performance and better capture the semantic relationships between genes and phenotypes.
- **Biological Validation**: Interpret the clusters in the context of real-world biological data to determine whether the clusters represent meaningful biological groupings.