Network inference in sparse datasets

Student Name: Harsh Pandey
Roll Number: 2021462

BTP report submitted in partial fulfillment of the requirements
for the Degree of B.Tech. in Computer Science & Biosciences

on 28th November, 2023

BTP Track: Research
BTP Advisor
Vibhor Kumar

Indraprastha Institute of Information Technology
New Delhi

Student's Declaration

I hereby declare that the work presented in the report entitled **"Network inference in sparse datasets using machine learning and python"** submitted by me for the partial fulfillment of the requirements for the degree of *Bachelor of Technology* in *Computer Science & Biosciences* at Indraprastha Institute of Information Technology, Delhi, is an authentic record of my work carried out under guidance of **Prof. Vibhor Kumar** . Due acknowledgements have been given in the report to all material used. This work has not been submitted anywhere else for the reward of any other degree.

Student's Name : Harsh Pandey **Place & Date: New Delhi,29/11/2023**

Certificate

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

Advisors' Name: Prof. Vibhor Kumar **Place & Date: New Delhi,29/11/2023**

Abstract

This research addresses the challenge of network inference in large-scale, sparse datasets. Inspired by the Gini coefficient, we introduce a novel methodology leveraging machine learning, including the GENIE3 algorithm and Random Forest, for efficient network inference. Principal Component Analysis (PCA) is employed for dimension reduction, with systematic evaluations across varied sparseness levels. Notably, our study incorporates an in-house developed back-projecting methodology, enhancing the preservation of feature importances during the inference process. The adapted Gini-based algorithm, coupled with the proprietary backprojecting methodology, showcases promising results in sparse network inference and comparative analyses with SVD and NMF. This research contributes to the acceleration and reliability of network inference, offering a valuable advancement in the field.

Keywords: Network Inference, Large-scale Networks, Sparse Datasets, Machine Learning, Random Forest, GENIE3 Algorithm, Gini Coefficient, Principal Component Analysis (PCA), Gene Network Inference, Matrix Factorization Techniques, SVD, NMF .

Acknowledgments

I would like to thank my project advisor, Vibhor kumar, for his continuous support, guidance, and motivation.I am deeply grateful for his guidance and for the lab facilities and instruments provided for the implementation of this project. Their assistance was invaluable in this project.

Work Distribution

The ideation phase of this project began in August and was followed by weekly discussions with Dr. Vibhor Kumar in september, aimed at refining the research idea. In October and November, we successfully developed and implemented our backprojecting methodology to tackle sparsity challenge offer by large dataset. During this time we run our model on more than 30+ different dataset from UCI ML to check its robustness in different datsets.Looking ahead, we continued to explore different research directions further to enhance the quality and realism of our model and to implement it on genomic data and categorical datsets finally.

Contents

Chapter 1

Introduction

1.1 Background

Network inference is a research approach widely applied in biology and computer science to uncover regulatory connections within complex systems. In biology, it helps unveil relationships between molecular components like genes or proteins, contributing to a deeper understanding of their interactions. Despite the availability of various methods for this purpose, there remains a need to explore the statistical formulations, connections, and differences between these approaches.

In computer science, network inference extends to improving the overall user experience by enabling the network end-system (computer) to deduce properties about network and end-system behavior. This leads to enhancements such as optimized resource sharing and reduced latency through minimized queuing. Network inference is not confined to a single domain; its applications span diverse fields, accommodating a broad range of data sources, including gene and genomic data.

Essentially, network inference acts as a crucial tool to decipher latent relationships within complex systems, offering valuable insights across scientific and technological domains. Its continuous refinement plays a pivotal role in advancing our comprehension of intricate networks and their impact on various facets of our world.

1.2 Motivation and Research Problem

As the interconnectedness of nodes in networks grows exponentially, the task of inferring meaningful connections becomes increasingly arduous. Large-scale networks, particularly those characterized by sparse datasets, present a considerable computational challenge for traditional inference methodologies. Motivated by the critical need for accelerated and efficient network inference, this research embarks on a journey to bridge the gap between computational limitations and the demands of complex systems.

The motivation for this study is underscored by the recognition that current methodologies may fall short in providing timely and accurate insights into networks characterized by numerous nodes and sparse data. Drawing inspiration from the ubiquity and efficiency of the Gini coefficient in statistical analysis, our research seeks to adapt and enhance its principles for robust network inference.

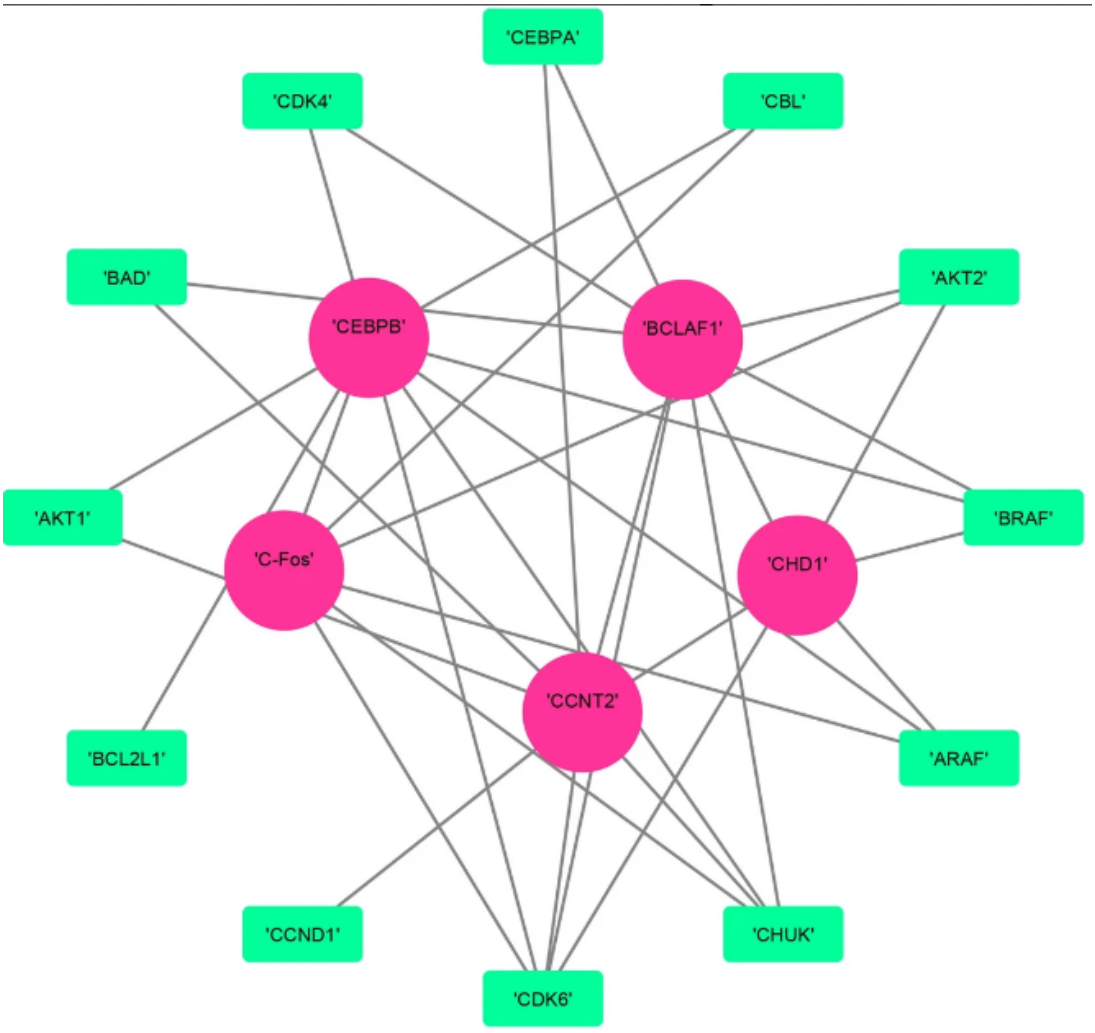

Figure 1.1

Amidst these considerations, the research problem crystallizes: How can we develop a methodology that not only expedites network inference in the face of sparse datasets but also sets the stage for meaningful insights into the intricate networks governing diverse domains? The exploration begins with an analysis of sparse datasets from the UCI Machine Learning Repository, setting the foundation for a methodology that not only overcomes computational hurdles but also anticipates broader applications, notably in the realm of genomic data.

Chapter 2

Research Approach and Work

2.1 How to infer networks from large datasets

1. **Data pre-processing**: Publicly available datasets can have noise and redundant information which can be a problem in our network inference. To address this, we need to remove these issues using a variety of techniques for normalization, such as standard scaler, qq norm, etc.

2. **Selecting an appropriate inference method**: There are a variety of methods available for network inference, and based on our requirements, we can choose the most suitable one. For data with noise and high dimensions, machine learning methods like random forest classifier or regressor can be used. In our project, we have made use of a random forest regressor.

3. **Parameter selection**: Many inference methods have parameters that need to be set. The optimal values of these parameters depend on the data and the inference method used. Thus, it is important to carefully select the parameter values and assess their impact on the results.

4. **Network validation**: Once a network is inferred, it is crucial to validate it using additional data or experimental validation. This helps to ensure that the network accurately reflects the underlying biological or social processes.

2.2 Algorithms for inferring network

1. **Genie3**: GENIE3 (Gene Network Inference with Ensemble of trees) it is an algorithm for inferring networks which makes use of the decision tree ensembles to corelate the corelation and information between the various genes available in the network . On advantage of this is that it has the power to infer both directed as well as undirected networks . The basic idea behind GENIE3 is to use random forests, which are an ensemble of decision trees, to predict the expression of a target gene based on the expression of other genes in the dataset. The importance of each predictor gene in the random forest is then used as a measure of the strength of the regulatory interaction between the predictor and the target gene. This process is repeated for each gene in the dataset, resulting in a complete regulatory network. One main disadvantage of genie 3 is an assumption that which assumes that the interactions between the entries are in a linear manner . It also does not consider

the dependencies of temporal type in the network . It is the most widely used method for network inference .

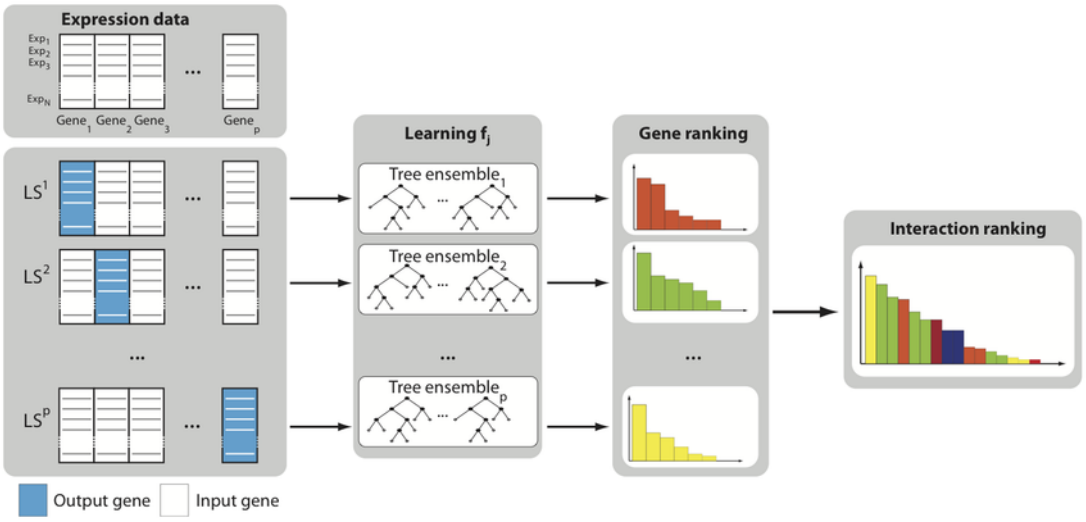

Figure 2.1

2. **Random Forest for network inference**: Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. Random forest is a type of classifier which makes the use of multiple decision trees and takes the decision of the majority of the trees decision and tries to make a prediction on the dataset . It is of two types classifier and regressor . It can be used to infer the network by calculation of the feature importance's of the target variables and then try to form a relationship between them to infer the network . The basic idea is to use random forest to predict the expression level of a target gene based on the expression levels of other genes in the dataset. The importance scores of the predictor genes are then used as a measure of the strength of the regulatory interaction between the predictor and the target gene.
To apply random forest for network inference, the following steps can be taken:

i) Data pre-processing: Here the data is is normalized and cleaned using the methods discussed above .
ii) Training the random forest: We then train our random forest model using the target as our y and our data minus y as x and then we predict using x and y .
iii) Determining variable importance: The variable importance measures are calculated to determine the strength of the regulatory interactions between the predictor and target genes. The importance measures can be calculated using different methods, such as mean decrease impurity or mean decrease accuracy.
iv) Network construction: Once the variable importance measures are obtained, a regulatory network can be constructed by connecting the predictor genes to the target gene using edges that correspond to the strength of the regulatory interaction.
v) Network validation: The network should be validated using external datasets or experimental validation methods to ensure that it accurately reflects the underlying biological or social processes. We can also use random forest along with other ml methods like neural networks or gradient boost to improve performance .

2.3 Our Methodology

```python
def corell(pcimp, i1, X, x_pca, met):
    r = X.shape[0]
    corr_matrix = np.zeros((r, i1))

    for i in range(X.shape[1]):
        for j in range(i1):

            if np.std(X.iloc[:, i]) == 0.0 or np.std(x_pca[:, j]) == 0.0:
                corr = 0.0
            else:
                corr, _ = spearmanr(X.iloc[:, i], x_pca[:, j])
            corr_matrix[i][j] = abs(corr)

    res = [0] * X.shape[1]

    if met == 1:
        for ii in range(X.shape[1]):
            if np.std(X.iloc[:, ii]) == 0.0:
                res[ii] = 0.0
            else:
                feat1 = 0
                for j in range(i1):
                    feat1 = feat1 + pcimp[j] * corr_matrix[ii][j]
                temp = 0
                for k in range(i1):
                    temp = temp + corr_matrix[ii][k]
                res[ii] = feat1 / temp

    return res
```

Figure 2.2: Our Backprojecting Methodology

2.3.1 Backprojecting Methodology Explanation

Explanation:

1. **Input Parameters:**
   - `pcimp`: Feature importances obtained from the RandomForest classifier.
   - `i1`: Number of components used in the NMF (Non-Negative Matrix Factorization).
   - `X`: Original feature dataset.
   - `x_pca`: Transformed dataset obtained from NMF.
   - `met`: Method indicator (1 or 2).
2. **Computing Correlation Matrix:** The function initializes a correlation matrix (`corr_matrix`) with zeros to store correlations between features and NMF components. It iterates over each feature in the original dataset (`X`) and each component in the NMF-transformed

dataset (`x_pca`). For each pair of feature and component, it calculates the Spearman correlation coefficient (`corr`) and stores its absolute value in the `corr_matrix`.

$$\rho = \frac{cov(rank(X), rank(Y))}{\sqrt{var(rank(X)) \times var(rank(Y))}}$$

3. **Backprojection Calculation:** The function initializes an array `res` to store the backprojected values for each feature. It then proceeds based on the specified `met` method:

   - **Method 1 (`met == 1`):** For each feature, it calculates a weighted sum (`feat1`) of NMF components using the provided feature importances (`pcimp`) and correlation matrix entries. It calculates the sum of correlation matrix entries for normalization (`temp`). The backprojected value for each feature is computed as the ratio of `feat1` to `temp`.

$$WeightedSum(\texttt{feat1}) = \sum_{j=1}^{i1} \texttt{pcimp}[j] \times \texttt{corr\_matrix}[featureindex][j]$$

4. **Returning Backprojected Values:** The function returns an array (`res`) containing the backprojected values for each feature.

Chapter 3

Results

3.1 Overview

In this section, we delve into the outcomes of our network inference methodology applied to diverse datasets. The research journey involved crucial steps, starting with the computation of feature importances through a Random Forest Regressor. Subsequently, we explored the impact of dimension reduction techniques such as PCA and NMF on the dataset. Our analysis included backprojecting features and assessing the correlation between feature importances in sparse-induced datasets.

3.2 Data Description

To conduct a comprehensive evaluation, we utilized 30+ datasets from UCI ML , each representative of distinct network structures,to check robustness of our methodology. The datasets underwent preprocessing, ensuring removal of noise and redundancy to enhance the quality of network inference.

3.3 Discussion of Findings

Our findings indicate a notable outcome: following backprojection on datasets, there was minimal degradation in feature importances. Particularly in scenarios of high sparseness, our backprojection model performance in both "back normal" and "back sparse" outperforms the normal results. Our model not only expedites the inference process but also preserves the integrity of feature importances. This underscores the effectiveness of our approach in enhancing the efficiency of network inference without compromising on quality.

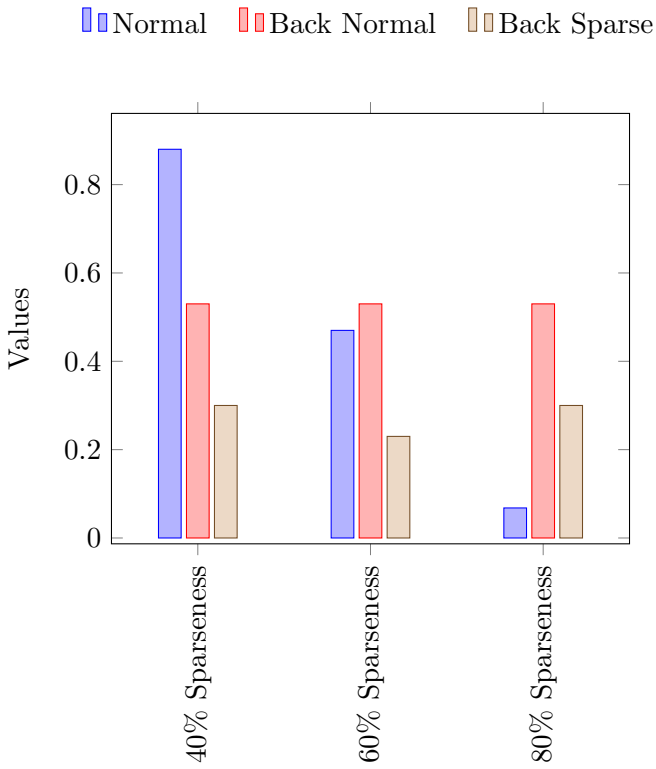

Figure 3.1: Comparison of Values at Different Sparseness Levels

	Sparseness	Components	Normal	BackNormal	BackSparse	Other
0	10.0	2.0	0.915520	0.150239	0.352818	0.661959
1	10.0	6.0	0.915520	0.255895	0.121500	0.410137
2	10.0	10.0	0.915520	0.340757	0.130771	0.514107
3	10.0	14.0	0.915520	0.409842	0.130503	0.376630
4	10.0	18.0	0.915520	0.533720	0.154395	0.262737
5	20.0	2.0	0.892208	0.150229	0.346420	0.656848
6	20.0	6.0	0.892208	0.255704	0.121176	0.407640
7	20.0	10.0	0.892208	0.340862	0.130893	0.524006
8	20.0	14.0	0.892208	0.410149	0.129070	0.389205
9	20.0	18.0	0.892208	0.535723	0.150615	0.265006
10	40.0	2.0	0.886262	0.150232	0.303083	0.683308
11	40.0	6.0	0.886262	0.255648	0.124924	0.444071
12	40.0	10.0	0.886262	0.340944	0.134635	0.540144
13	40.0	14.0	0.886262	0.409857	0.135967	0.423618
14	40.0	18.0	0.886262	0.534317	0.165374	0.310455
15	60.0	2.0	0.473806	0.150235	0.266990	0.764012
16	60.0	6.0	0.473806	0.255658	0.128883	0.478947
17	60.0	10.0	0.473806	0.340696	0.106229	0.567209
18	60.0	14.0	0.473806	0.410201	0.136974	0.520307
19	60.0	18.0	0.473806	0.534727	0.233616	0.497780
20	80.0	2.0	0.068219	0.150236	0.302861	0.490826
21	80.0	6.0	0.068219	0.255530	0.103014	0.398914
22	80.0	10.0	0.068219	0.340823	0.115893	0.514962
23	80.0	14.0	0.068219	0.409925	0.145812	0.433821
24	80.0	18.0	0.068219	0.534866	0.136157	0.218345

Figure 3.2

Bibliography

[1] Simon Knott and others. *Genetic Network Inference via Gene Set Stochastic Sampling and Sensitivity Analysis*. In *Proceedings of the IEEE International Conference on Control Applications*, 2005. 10.1109/cca.2005.1507116

[2] Mohamed Abbas and others. *Mixed Machine Learning Approach for Efficient Prediction of Human Heart Disease by Identifying the Numerical and Categorical Features*. In *Applied Sciences*, volume 12, number 15, 2022, pages 7449.

[3] Authors not available. *A systematic evaluation of single-cell RNA-sequencing imputation methods*. In *Genome Biology*, volume 22, number 1, 2020, pages 33. `https://genomebiology.biomedcentral.com/articles/10.1186/s13059-020-02132-x`

[4] *GENIE3 Procedure*. `https://www.researchgate.net/figure/GENIE3-procedure-For-each-gene-a-learning-sample-is-generated-with-expression-levels_fig12_47357972`