Project Proposal

Project Title: Thresholded Boolean Co-Abundance (ESABO) analysis for continuous-valued datasets
Student Name: Devi Chandran
Student ID: 2430977
Supervisor Name: Jens Christian Claussen

Project Category/Topic: Data Science

Project Aim: The aim of the project is to implement Boolean co-abundance analysis to continuous-valued data, use binarization(or thresholding) to convert to binary data(presence or absence data, represented as binary values, i.e., (1 or 0)) and optimize the threshold for maximal information gain. The maximal information gain is attained using Entropy Shifts of Abundance Vectors under Boolean Operations(ESABO) method using the information gain from the pairs of abundance vectors(abundance of entities), when combined via Boolean operations(as described in [1]). The goal would be then to extract networks formed through the ESABO method on multivariate brain activity data and to analyse it using network complexity measures [2].

Significance:
Boolean co-abundance analysis is significant in finding the correlation or dependencies between variables in binary(or Boolean) data. The ESABO method used in our project has been proven to study complex biological datasets in binary framework and provide an idea of the systematic interactions or relationships in microbial ecosystems.

Relevance to data science:
Our project aims to incorporate the Boolean co-abundance analysis and its relevance to data science is as follows:

- **Correlation/Association**: Use the ESABO method for analysing binary data to find the different correlation or dependencies between a number of variables
- **Feature selection**: The property of selecting the most significant associations of variables allows us to define a feature set/group. This indeed is beneficial in machine learning algorithms for dimensionality reduction.
- **Network analysis**: The Boolean co-abundance analysis can be beneficial in creating the networks or graphs of such associations(nodes representing variables, links representing the connections between the nodes) and using network science methods to get a deeper understanding of the data structure.

Related work:

The paper[1] introduces the ESABO method for inferring microbial interaction networks from abundance data and tests the same using simulation data. The obtained method is applied to a new data set of human gut microbiome compositions to show the statistically significant interactions in the co-abundance networks among low-abundance species. In the paper [3], the concept of information theory and entropy is well explained and can be of high significance in our project.

Project Objectives/Deliverables:

1. **Objective :** Implementing ESABO and testing it on a simple dataset
   Deliverable: The ESABO method can be tested on a simple binary dataset(preferably a microbiome dataset) to ensure its working and the report on “entropy shift”(through

calculated entropies), i.e., z-score(to calculate the positive or negative interactions between the species of microbiomes) will be presented.

2. **Objective** : For continuous-valued data, use binarization (thresholding) to convert to binary data:
3. **Deliverable:** The binary data would be presented as a proof of thresholding to the supervisor.

4. **Objective** : Optimize the threshold for maximal information gain
   Deliverable: ESABO scores would be systematically calculated depending on the threshold value.

5. **Objective** : Extract networks via ESABO and simple correlation analysis and comparison of both
   Deliverable: The results of both the methods would be documented, these would include the graphs obtained from the networks of both methods(links and interactions).

6. **Objective** : Add network complexity measures(and other network science methods) to analyse the networks obtained
   Deliverable: Degree distribution and other network graphs

The above objectives are sufficient to attain the project aim as it has concrete steps to be done at various stages of the project. Successful completion of each of the steps at suitable time steps will make it easier to reach the goal.

Methodology:

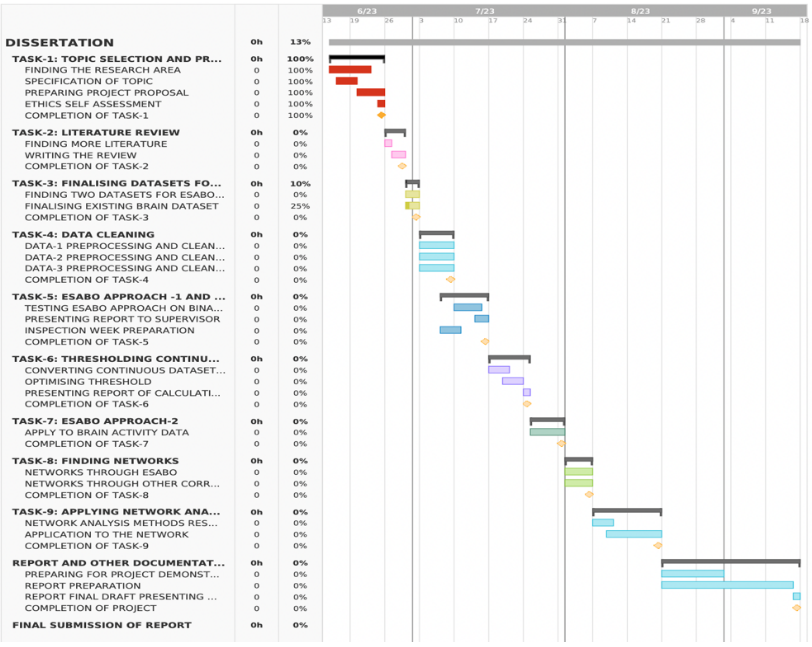
- **Boolean co-abundance analysis using ESABO method:**Our project uses the ESABO method for evaluating the information content of abundance binary data and find the co-abundance patterns for indication of highly significant features. It assigns to each abundance a Boolean operation and evaluates the z-score of the entropy of abundances. The method is applied first to a simple dataset and then to a continuous valued dataset(after thresholding) to attain the optimal threshold value.The method would be then tested at a later stage to a multivariate brain activity data which is appropriate to show that our method(ESABO) works on continuous real valued data.

- **Network complexity measures:** Later stages of the project would include the analysis of the network formed out of ESABO and other correlation analysis method. Both the networks would be compared and contrasted to obtain the optimal network for our problem and the same would be analysed using multiple network complexity measures as in [2].

Project plan:

Feasibility: My expertise in machine learning algorithms, data science and mathematics would be highly beneficial in attaining the goal. I am genuinely interested in the project as it would align greatly to my interest in mathematics and computational biology. My current postgraduate degree in data science and past undergraduate in mathematics would definitely be applicable in this dissertation. I believe that, under guidance of my supervisor(who is an expert in this field having applied the methodology to his previous researches and himself suggesting the topic) I would be able to complete my project within due time.

Gantt chart with tasks and milestones, reflecting the project objectives/deliverables



DISSERTATION	**0h**	**13%**
TASK-1: TOPIC SELECTION AND PR...	**0h**	**100%**
FINDING THE RESEARCH AREA	0	100%
SPECIFICATION OF TOPIC	0	100%
PREPARING PROJECT PROPOSAL	0	100%
ETHICS SELF ASSESSMENT	0	100%
COMPLETION OF TASK-1	0	100%
TASK-2: LITERATURE REVIEW	**0h**	**0%**
FINDING MORE LITERATURE	0	0%
WRITING THE REVIEW	0	0%
COMPLETION OF TASK-2	0	0%
TASK-3: FINALISING DATASETS FO...	**0h**	**10%**
FINDING TWO DATASETS FOR ESABO...	0	0%
FINALISING EXISTING BRAIN DATASET	0	25%
COMPLETION OF TASK-3	0	0%
TASK-4: DATA CLEANING	**0h**	**0%**
DATA-1 PREPROCESSING AND CLEAN...	0	0%
DATA-2 PREPROCESSING AND CLEAN...	0	0%
DATA-3 PREPROCESSING AND CLEAN...	0	0%
COMPLETION OF TASK-4	0	0%
TASK-5: ESABO APPROACH -1 AND ...	**0h**	**0%**
TESTING ESABO APPROACH ON BINA...	0	0%
PRESENTING REPORT TO SUPERVISOR	0	0%
INSPECTION WEEK PREPARATION	0	0%
COMPLETION OF TASK-5	0	0%
TASK-6: THRESHOLDING CONTINU...	**0h**	**0%**
CONVERTING CONTINUOUS DATASET...	0	0%
OPTIMISING THRESHOLD	0	0%
PRESENTING REPORT OF CALCULATI...	0	0%
COMPLETION OF TASK-6	0	0%
TASK-7: ESABO APPROACH-2	**0h**	**0%**
APPLY TO BRAIN ACTIVITY DATA	0	0%
COMPLETION OF TASK-7	0	0%
TASK-8: FINDING NETWORKS	**0h**	**0%**
NETWORKS THROUGH ESABO	0	0%
NETWORKS THROUGH OTHER CORR...	0	0%
COMPLETION OF TASK-8	0	0%
TASK-9: APPLYING NETWORK ANA...	**0h**	**0%**
NETWORK ANALYSIS METHODS RES...	0	0%
APPLICATION TO THE NETWORK	0	0%
COMPLETION OF TASK-9	0	0%
REPORT AND OTHER DOCUMENTAT...	**0h**	**0%**
PREPARING FOR PROJECT DEMONST...	0	0%
REPORT PREPARATION	0	0%
REPORT FINAL DRAFT PRESENTING ...	0	0%
COMPLETION OF PROJECT	0	0%
FINAL SUBMISSION OF REPORT	**0h**	**0%**

Explanation of Gantt chart : The project starts from 26-06-2023 as soon as the proposal is approved. The entire project is broken down into several tasks and accompanying milestones at the completion of each task. The project begins with finding more specific research goals through a refined literature review of research articles associated with project. The same week would include finding appropriate datasets for the project and would be completed by 01-07-2023. The next week(week 4) would involve the pre-processing and cleaning of the datasets for the application of the method. This week for also involve preparation for the inspection week with appropriate documentation. Week 5(10-07-2023) would involve project inspection and application of ESABO method to a sample dataset. The report of the same would be presented to the supervisor.

The main part of the project starts in Week-6(17-07-2023) and would involve testing the approach on continuous valued datasets, thresholding. The next week, Week-7(24-07-2023) would be application of ESABO approach to brain activity data and presenting of the results to the supervisor. The same would be then used to find the networks in week 8(01-08-2023) and analysis of the network in week 9(07-08-2023). Week 9 and 10 would be highly occupied with the analysis and results, and will form another major part of the project. These weeks would be also comprising of checking the results and solutions arrived at. Finally, the following weeks would be dedicated to report writing and other documentation for final submission. The project is expected to finish by 14th of September.

Risks and contingency plan:

- Datasets required for the project are in raw format and may require more time for pre-processing and loading. Plan is to seek help from technical support provided during the weekdays, along with guidance from supervisor and proper self-training.
- Network analysis methods would be a bit risky to perform on datasets and the same could be tackled with more reading and coding help.

- The project is highly mathematical and would require more input from my side to understand and produce code for it. I will seek help from existing coding techniques on GitHub and other platforms for the same
- Another task is to understand the medical data, i.e. brain activity data in our project and to tackle this I will use the help of medical practitioners and neuroscience experts.

Hardware/Software Resources

The project would not involve any amount of hardware/software resources other than python programming.

Data

The datasets required in the project are available in open databases such as the (multivariate) brain activity data, microbial and ecological data. There would be no issues in accessing the data as it is openly accessible to the public for use.

References

[1] "Boolean analysis reveals systematic interactions among low-abundance species in the human gut microbiome(2017)" by Jens Christian Claussen , Jurgita Skiecevičienė, Jun Wang, Philipp Rausch, Tom H. Karlsen, Wolfgang Lieb, John F. Baines, Andre Franke, Marc-Thorsten Hütt.- https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005361

[2] **"**Network Science**"** by Albert-László Barabási.- http://networksciencebook.com

[3] "The Mathematical Theory of Communication"(University of Illinois Press, Urbana, 1949) by C. E. Shannon and W. Weaver-
https://pure.mpg.de/rest/items/item_2383164/component/file_2383163/content