Bioinformatics exam project: Disease subtype discovery using multi-omics data integration

Jessica Gliozzo, Giorgio Valentini

Contents

General organization of the exam

The exam is composed by two parts:

1. The student will prepare a software project (see details in section Project) comprising:
   a. The runnable code to replicate the results of the project. This can be a simple R script or a R notebook (in this last case, please provide the R Markdown .Rmd file). Please provide also the output of the command *sessionInfo()* that provides a report with version information of R and used packages. The code should be appropriately commented in order to evaluate if the student understood the code.
   b. A report describing the considered problem, the machine learning approaches applied to solve it, the experimental set-up, the results obtained and a discussion of the results. The report comprises common sections present in a scientific paper (see section Report for further details). The code and report of the project must be delivered by e-mail **at least 7 days** before the exam. Questions about the project can be sent via e-mail (jessica.gliozzo@unimi.it).
2. There will be an **oral exam**. The oral exam will be a discussion about possible errors present in the project (code and report), general questions about the project and it will also evaluate the student knowledge about all the topics explained during the lessons.

Project

The project can be done alone or in groups (up to three students per group). The project will be lighter if done alone, while for groups additional mandatory parts are requested. Mandatory parts for groups are marked with the tag [GROUP]. If you are alone but you want to do all the parts, feel free to proceed. When working in groups, each student should contribute equally to the writing of both code and report and **any member of the group should have understood the project in its entirety** (i.e. not just the parts wrote by the single student). There is a task tagged as [OPTIONAL], which is optional for both groups and single students (but in for a penny, in for a pound!).

The project regards the discovery of disease subtypes using a multi-omics dataset coming from TCGA. The dataset is the Prostate adenocarcinoma dataset (disease code "PRAD"). We will consider as disease subtypes the ones identified in a work performed by *The Cancer Genome Atlas Research Network* [1], where they used

an integrative clustering model (called iCluster [2]) on multi-omics data (somatic copy-number alterations, methylation, mRNA, microRNA, and protein levels) and discovered three disease subtypes. The student(s) should:

1. Download the Prostate adenocarcinoma dataset considering three different omics data sources (mRNA, miRNA and protein expression data). The TCGA code for the dataset is "PRAD".
2. Pre-process the dataset following the same steps we used during lessons. During the filtering by variance, select the first 100 features having highest variance from each data source.
3. Download the disease subtypes (column "Subtype_Integrative" is the one containing the iCluster molecular subtypes). Note that not all subtypes are available for the set of samples having all the considered omics data sources, thus you need to retain from the multi-omics dataset only samples having an associated subtype.
4. Check that patients in multi-omics dataset and subtypes are in the same order.
5. Integrate the data using Similarity Network Fusion [3] with the scaled exponential euclidean distance.
6. Try to integrate the similarity matrices from each data source (computed by scaled exponential euclidean distance) using a simple average of the matrices. This can be considered as a trivial multi-omics data integration strategy.
7. [GROUP] Integrate the dataset using another data fusion method called NEMO [4] to obtain an integrated similarity matrix. NEMO implementation is available on github (https://github.com/Shamir-Lab/NEMO).
8. Perform disease subtype discovery (number of clusters equal to the number of disease subtypes found by iCluster) using PAM algorithm [5] on the following similarity matrices:
    a. Similarity matrices obtained from single data sources (i.e. miRNA, mRNA, proteins) using the usual scaled exponential euclidean distance. Thus, you should obtain three different similarity matrices. To compute the corresponding distance matrix use this code: dist <- 1 - NetPreProc::Max.Min.norm(W). Max.Min.norm() function is in the NetPreProc CRAN package (https://cran.r-project.org/web/packages/NetPreProc/index.html). The idea is to normalize the similarity matrix before computing the corresponding distance.
    b. Integrated matrix obtained using the average among matrices. Use dist <- 1 - NetPreProc::Max.Min.norm(W) to compute the distance matrix.
    c. Integrated matrix obtained using Similarity Network Fusion. Use dist <- 1 - NetPreProc::Max.Min.norm(W) to compute the distance matrix.
    d. [GROUP] Integrated matrix obtained using NEMO. Use dist <- 1 - NetPreProc::Max.Min.norm(W) to compute the distance matrix.
9. [GROUP] NEMO provides the possibility of performing clustering using another approach called Spectral Clustering [6]. Use the function *nemo.clustering()* to test this approach.
10. [OPTIONAL] Apply Spectral Clustering on the integrated matrix obtained using Similarity Network Fusion (an implementation of spectral clustering is *SNFtool::spectralClustering()*, which is the same exploited in *nemo.clustering()*).
11. Compare the clusterings obtained by each considered approach w.r.t. the iCluster disease subtypes. Make tables and plots to show the results and discuss them.

Set the neighborhood size to 20 (k=20) for scaled exponential euclidean distance, SNF and NEMO.

Programming language: Students can use R or Python to write the code of the project. Since the course was delivered using only R, students that decide to use Python need to find the appropriate implementations for each method necessary to comply with the requests of the project. No support in this sense will be provided.

Report

The report must contain the following sections:

1. INTRODUCTION: Illustrate the problem of disease subtype discovery from multi-omics data (explore the literature about this topic), considering the problems of data integration and clusters computation.
2. METHODS: describe the data integration and clustering approaches exploited. Describe also used dataset, considered disease subtypes and the metrics employed to compare the obtained clusterings with the disease subtypes. Explain also the data-preprocessing applied.
3. RESULTS: Present the results of the various approaches using tables and plots. Discuss the obtained results. Note that the results do not necessarily need to be good.

References

[1] A. Abeshouse *et al.*, "The molecular taxonomy of primary prostate cancer," *Cell*, vol. 163, no. 4, pp. 1011–1025, 2015.

[2] R. Shen, A. B. Olshen, and M. Ladanyi, "Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis," *Bioinformatics*, vol. 25, no. 22, pp. 2906–2912, 2009.

[3] B. Wang *et al.*, "Similarity network fusion for aggregating data types on a genomic scale," *Nature methods*, vol. 11, no. 3, pp. 333–337, 2014.

[4] N. Rappoport and R. Shamir, "NEMO: Cancer subtyping by integration of partial multi-omic data," *Bioinformatics*, vol. 35, no. 18, pp. 3348–3356, 2019.

[5] "Partitioning around medoids (program PAM)," in *Finding groups in data*, John Wiley & Sons, Ltd, 1990, pp. 68–125. doi: https://doi.org/10.1002/9780470316801.ch2.

[6] U. Von Luxburg, "A tutorial on spectral clustering," *Statistics and computing*, vol. 17, pp. 395–416, 2007.