BITS :: Call for Abstracts 2022 - Oral communication

Type	Oral communication
Session	Machine Learning in Bioinformatics
Title	Comparison of early integration approaches for cancer survival prediction
All Authors	Gnuva M(1), Gliozzo J(1,2,3), Paccanaro A(4, 5), Valentini G(1,2), Casiraghi E*(1,2) corresponding author: Casiraghi E

Affiliation

(1) AnacletoLab, Dipartimento di Informatica Giovanni degli Antoni, Università degli Studi di Milano, ITALY
(2) Laboratorio Nazionale InfoLife, CINI
(3) European Commission, Joint Research Centre (JRC), Ispra, Italy
(4) School of Applied Mathematics (EMAp), Fundação Getúlio Vargas, Rio de Janeiro Brazil,
(5) Department of Computer Science, Royal Holloway, University of London, Egham, TW20 0EX UK

Motivation

Recent advances in high-throughput technologies, data transmission, and data storage, have allowed the generation, acquisition, and storage of huge amounts of data from multiple sources (multimodal datasets). In bioinformatics, exploiting such datasets requires the development of data integration techniques able to discard redundant information while enhancing the information characterizing each source. Among the techniques proposed in literature, early integration methods [doi: 10.1089/10665270252935539] assume that samples lie in a latent space from which multiple source-views are generated by unknown projections. This results in multiple data-views expressed as separate source-specific spaces that are characterized by both an individual and a shared structure (variance), where the latter causes collinearities between data-blocks. Therefore, early methods estimate the embedding into a shared latent space by minimizing the redundancy between the input data-blocks, while maximizing their individual variability. The resulting integrated representation may improve the results of subsequent analysis, supervised learning (e.g. for patient outcome prediction), or unsupervised clustering (e.g. patient subtype identification).

Methods

We analyzed data from the TCGA dataset through several experimental pipelines where Hierarchical PCA (HPCA, doi: 10.1002/cem.811) and Integrative Non-negative Matrix Factorization (iNMF, doi: 10.1093/bioinformatics/btv544) are compared and integrated. HPCA applies two consecutive PCA steps. The first PCA step is applied on each data-block separately to obtain a lower-dimensional, normalized data-block representation where within-source redundancies are minimized. These lower-dimensional data-blocks are then concatenated and used as input to the second PCA step whose aim is to remove between-source redundancies while extracting salient information. iNMF is an extension of joint Non-negative Matrix Factorization (jNMF, doi: 10.1093/nar/gks725). jNMF solves the data integration problem by solving multiple NMF problems subject to a shared factor matrix, therefore neglecting the individual information characterizing each source. To overcome this limitation, iNMF estimates factor matrices composed of both a shared and an individual structure. Our experimental pipelines are shown in Fig.1 (top). Besides comparing HPCA and iNMF, we also experimented a combination of PCA, for dimensionality reduction, with iNMF for integrating the computed lower dimensional data (PCA-iNMF pipeline). To remove redundancies in between the individual and shared spaces identified by iNMF we added a final PCA to the PCA-iNMF pipeline, therefore essentially integrating HPCA and iNMF (PCA-iNMF-PCA pipeline).

Results

In our experiments we used multi-omics data from the TCGA Breast cancer dataset. Each patient is described by high-dimensional methylation, miRNA and mRNA profiles, CNV scores and a binary outcome variable (progression-free interval event) The comparative evaluation of the results was performed by visual analysis of the clusters obtained by UMAP [doi: 10.48550/arXiv.1802.03426], a data-dimensionality techniques exploiting a manifold learning approach that guarantees the preservation of the local and global manifold structures, and by the application of Random Forest classifiers (RF, doi: 10.1023/A:1010933404324) on the integrated dataset. Next, considering the data unbalance, we also tested the application of data resampling techniques (undersampling, oversampling, SMOTE - doi: 10.1613/jair.953) and of cost-sensitive-learning during the training phase. In our preliminary results (Fig. 1-bottom), the best performance was achieved by the PCA-iNMF-PCA pipeline.

Info

3 authors (Casiraghi E, Gliozzo J, and Valentini G) are members of the CINI infolife laboratory

filename	pipeline_EN2.png

Figure

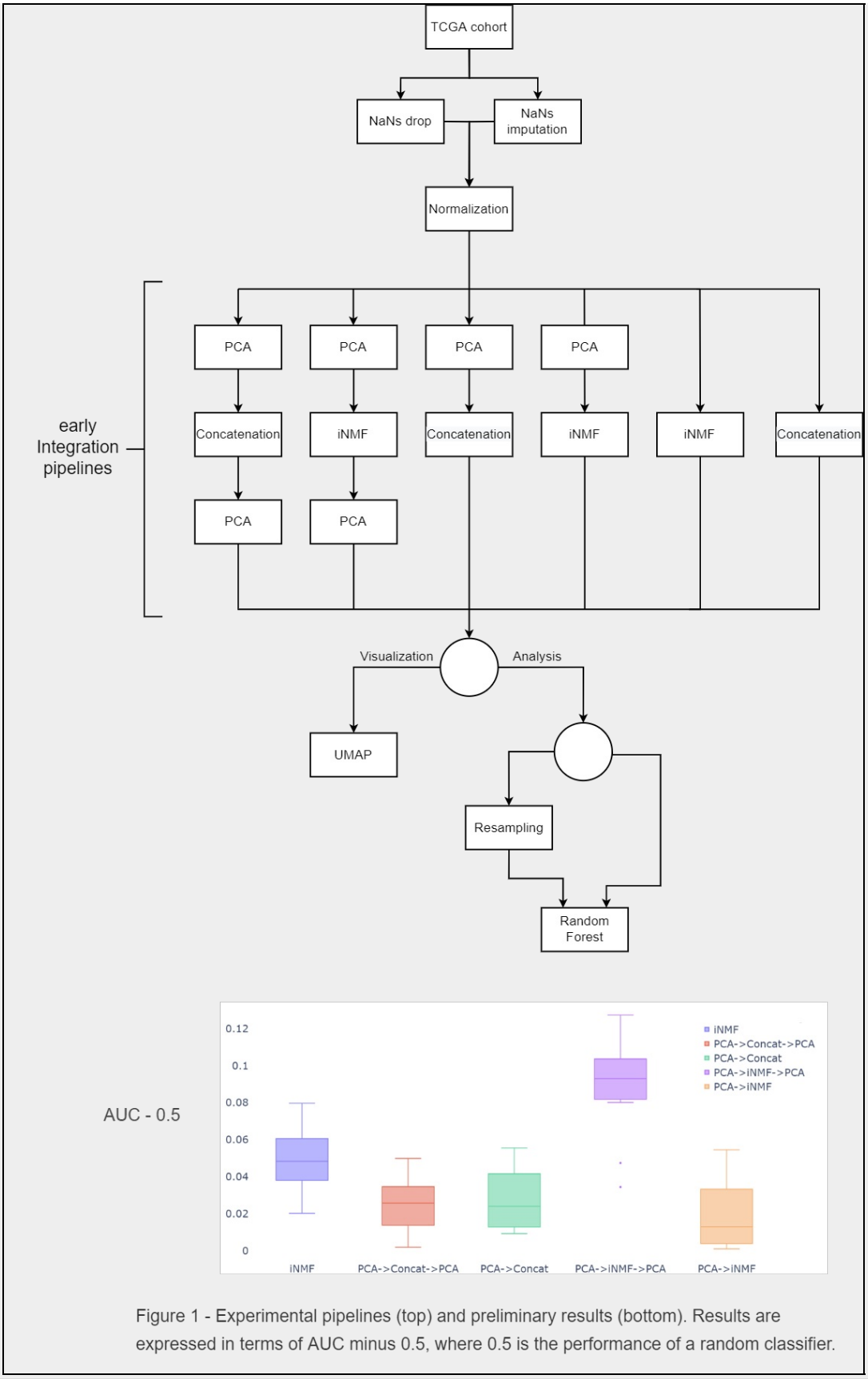

Figure 1 - Experimental pipelines (top) and preliminary results (bottom). Results are expressed in terms of AUC minus 0.5, where 0.5 is the performance of a random classifier.

Availability	https://github.com/mirco-gnuva/TCGA_tumore-seno_iNMF

Corresponding Author

Name, Surname	elena, casiraghi
Email	elena.casiraghi@unimi.it

Submitted on	27.04.2022