Multi-omics data integration methods: kernel and other machine learning approaches

Nathalie Vialaneix

nathalie.vialaneix@inrae.fr
http://www.nathalievialaneix.eu

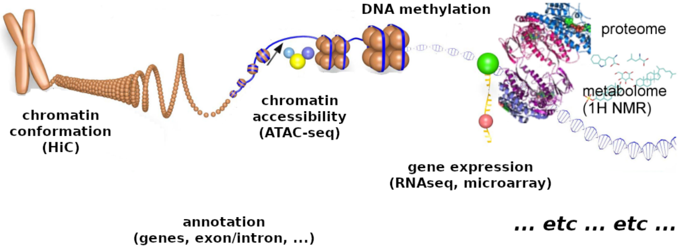
JOBIM 2023
27>30 JUIN
Édition multisite

Nice, June 29th 2023

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

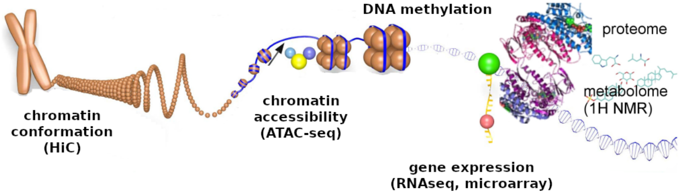
INRAE

Collected data at genomic level are increasingly publicly available


DNA methylation
proteome
chromatin accessibility (ATAC-seq)
metabolome (1H NMR)
chromatin conformation (HiC)
gene expression (RNAseq, microarray)
annotation (genes, exon/intron, ...)
... etc ... etc ...

the different levels are not always compatible

Collected data at genomic level are increasingly publicly available



DNA methylation

proteome

chromatin conformation (HiC)

chromatin accessibility (ATAC-seq)

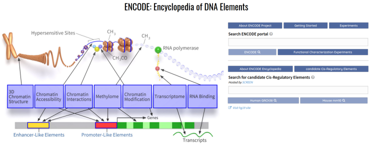
metabolome (1H NMR)

gene expression (RNAseq, microarray)

annotation (genes, exon/intron, ...)

... etc ... etc ...

the different levels are not always compatible



ENCODE: Encyclopedia of DNA Elements

TCGA

GTEx Portal

FAANG
Functional Annotation of Animal Genomes

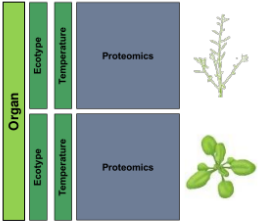
FR-AgENCODE

[Foissac et al., 2019]

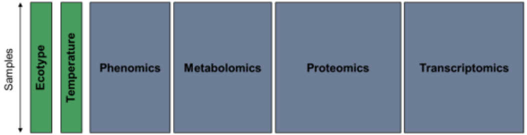
Omics data integration

Type of data to integrate

vertical:



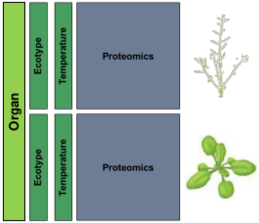
horizontal:



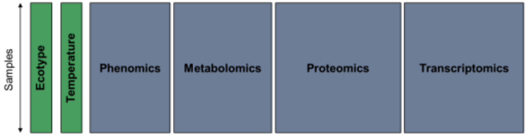
Omics data integration

Type of data to integrate

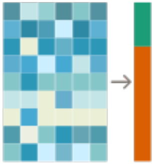
vertical:



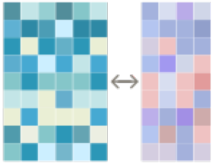
horizontal:



Type of analysis to perform



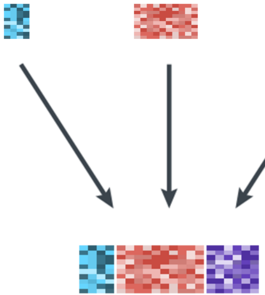
supervised:



unsupervised:

Left pictures courtesy Harold Duruflé

Multiple table analyses (CCA, MFA, PLS, STATIS, ...)



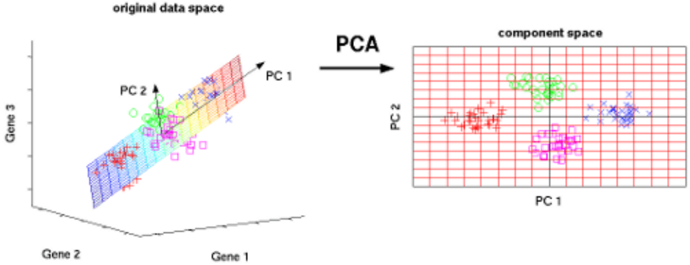
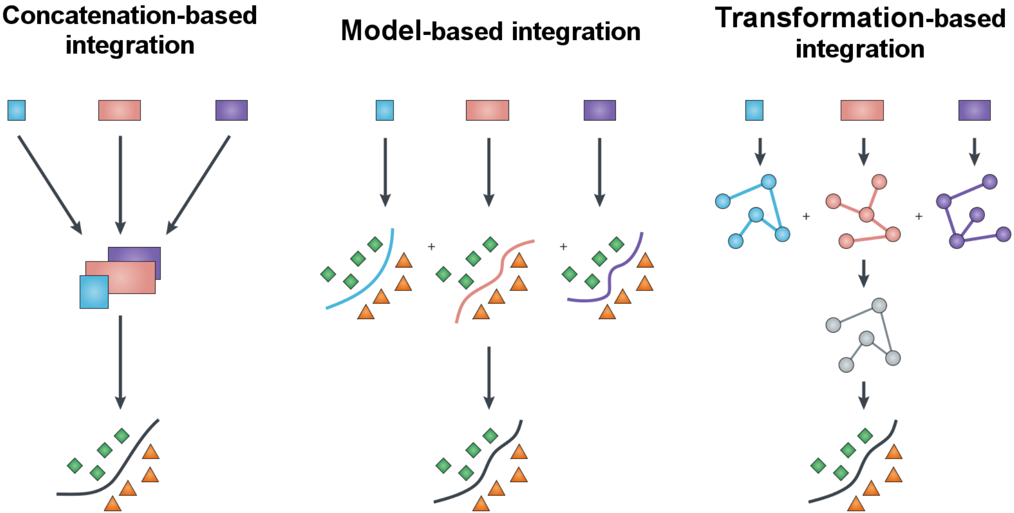


Image from https://dimensionless.in

Types of data integration methods [Ritchie et al., 2015]



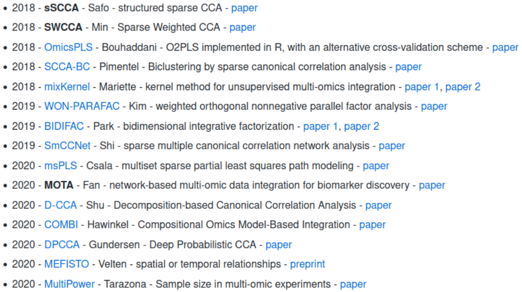
Concatenation-based integration

Model-based integration

Transformation-based integration

Want to know them all?

https://github.com/mikelove/awesome-multi-omics

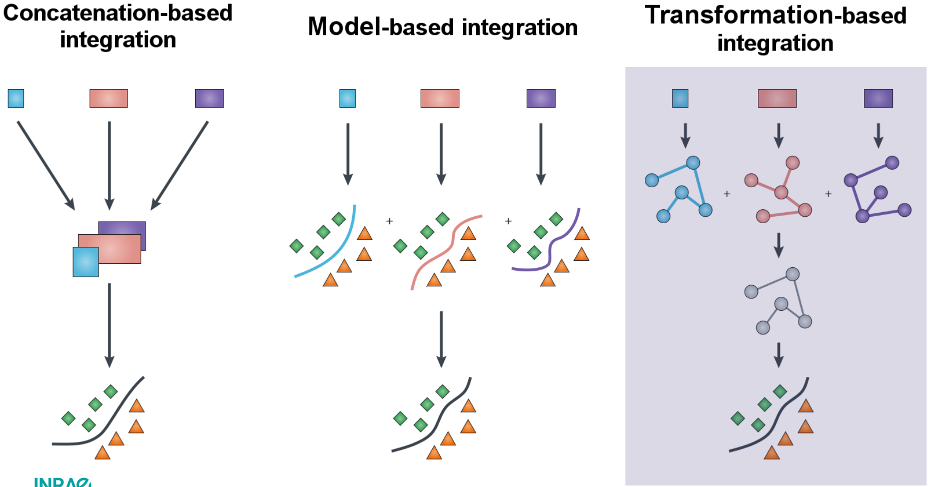


- 2018 - **sSCCA** - Safo - structured sparse CCA - paper
- 2018 - **SWCCA** - Min - Sparse Weighted CCA - paper
- 2018 - OmicsPLS - Bouhaddani - O2PLS implemented in R, with an alternative cross-validation scheme - paper
- 2018 - SCCA-BC - Pimentel - Biclustering by sparse canonical correlation analysis - paper
- 2018 - mixKernel - Mariette - kernel method for unsupervised multi-omics integration - paper 1, paper 2
- 2019 - WON-PARAFAC - Kim - weighted orthogonal nonnegative parallel factor analysis - paper
- 2019 - BIDIFAC - Park - bidimensional integrative factorization - paper 1, paper 2
- 2019 - SmCCNet - Shi - sparse multiple canonical correlation network analysis - paper
- 2020 - msPLS - Csala - multiset sparse partial least squares path modeling - paper
- 2020 - **MOTA** - Fan - network-based multi-omic data integration for biomarker discovery - paper
- 2020 - D-CCA - Shu - Decomposition-based Canonical Correlation Analysis - paper
- 2020 - COMBI - Hawinkel - Compositional Omics Model-Based Integration - paper
- 2020 - DPCCA - Gundersen - Deep Probabilistic CCA - paper
- 2020 - MEFISTO - Velten - spatial or temporal relationships - preprint
- 2020 - MultiPower - Tarazona - Sample size in multi-omic experiments - paper

Some specificities: can account for structure in data (network), are dedicated to a specific omic (single-cell), can account for temporal/spatial information, can include biological knowledge (mostly GO), ...

Scope of the rest of the talk

Unsupervised transformation based integration


Concatenation-based integration
Model-based integration
Transformation-based integration

Overview of the talk

Kernel methods

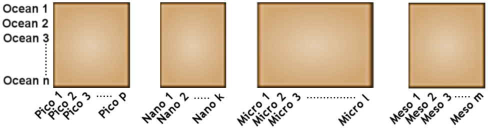
Integrating data with kernels

Conclusion, perspectives

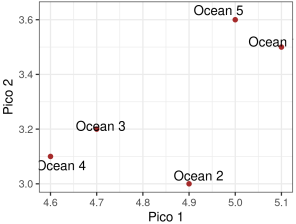
Main ideas behind kernel methods

Standard (omics) data analyses:

- data are (numeric) tables



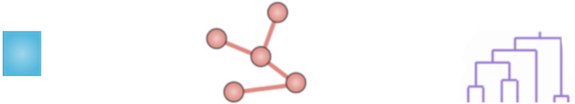
- analyses are based on operations (distances, means, ...) in the variable space



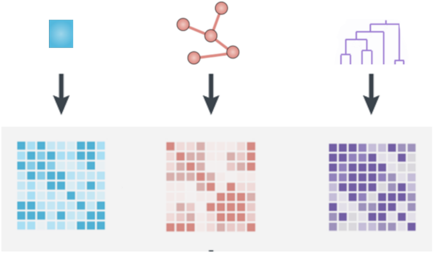
Main ideas behind kernel methods

Kernel data analyses:

- data are arbitrary



- analyses are based on transformations of data to “similarities” between samples

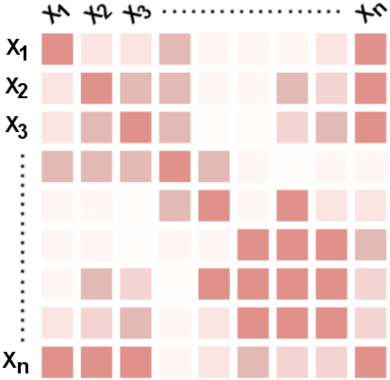


More formally...

n samples $(x_i)_i \in \mathcal{X}$

kernels: symmetric and positive $(n \times n)$-matrix

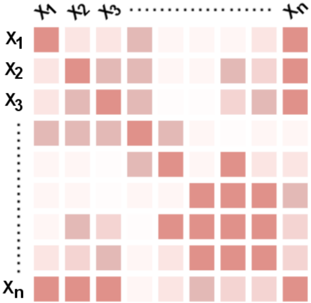
$\mathbf{K}$ that measures a "similarity" between n entities in $\mathcal{X}$



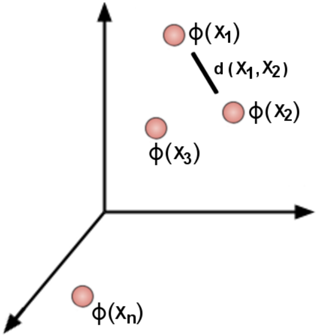
More formally...

n samples $(x_i)_i \in \mathcal{X}$

kernels: symmetric and positive $(n \times n)$-matrix **K** that measures a "similarity" between n entities in $\mathcal{X}$



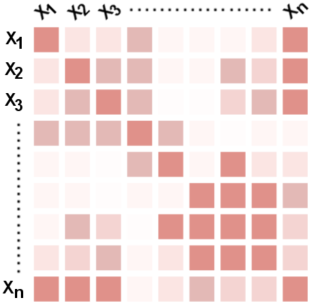
$$\mathcal{X} \xrightarrow{\Phi} \mathcal{H}$$



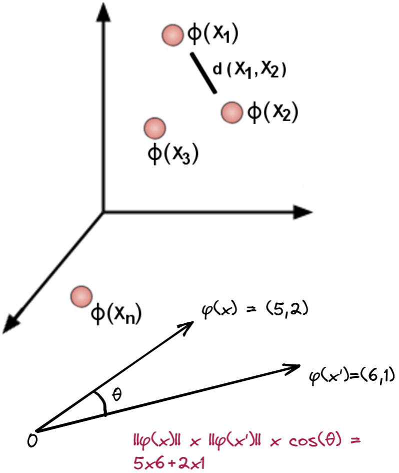
More formally...

n samples $(x_i)_i \in \mathcal{X}$

kernels: symmetric and positive $(n \times n)$-matrix **K** that measures a "similarity" between n entities in $\mathcal{X}$



$$\mathcal{X} \xrightarrow{\Phi} \mathcal{H}$$



$$\mathbf{K}(x, x') = \langle \phi(x), \phi(x') \rangle$$

Principles of learning from kernels

Start from any statistical method (PCA, regression, *k*-means clustering) and rewrite all quantities using:

- $\mathbf{K}$ to compute distances and dot products
  dot product is: $\mathbf{K}_{ii'}$ and distance is: $\sqrt{\mathbf{K}_{ii} + \mathbf{K}_{i'i'} - 2\mathbf{K}_{ii'}}$

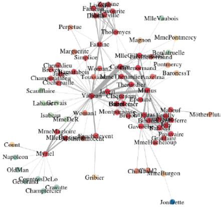
- (implicit) linear or convex combinations of $(\phi(x_i))_i$ to describe all unobserved elements (centers of gravity and so on...)

Kernel examples

1. $\mathbb{R}^p$ observations: Gaussian kernel $\mathbf{K}_{ii'} = e^{-\gamma \|x_i - x_{i'}\|^2}$

Kernel examples

1. $\mathbb{R}^p$ observations: Gaussian kernel $\mathbf{K}_{ii'} = e^{-\gamma\|x_i - x_{i'}\|^2}$



2. nodes of a graph: **[Kondor and Lafferty, 2002]**
3. sequence kernels (between proteins: spectrum kernel **[Jaakkola et al., 2000]** or convolution kernel **[Saigo et al., 2004]**)
4. kernel between graphs (used between metabolites based on their fragmentation trees): **[Shen et al., 2014, Brouard et al., 2016]**
5. kernel embedding phylogeny information for metagenomics **[Mariette and Villa-Vialaneix, 2018]**
6. ...

Overview of the talk

Kernel methods

Integrating data with kernels

Conclusion, perspectives

Multiple kernel (or distance) integration

How to “optimally” combine several kernel datasets?

For kernels $\mathbf{K}^1, \ldots, \mathbf{K}^M$ obtained on the same n objects, search: $\mathbf{K}_\beta = \sum_{m=1}^M \beta_m \mathbf{K}^m$ with $\beta_m \geq 0$ and $\sum_m \beta_m = 1$

Multiple kernel (or distance) integration

How to "optimally" combine several kernel datasets?

For kernels $\mathbf{K}^1, \ldots, \mathbf{K}^M$ obtained on the same n objects, search: $\mathbf{K}_\beta = \sum_{m=1}^M \beta_m \mathbf{K}^m$ with $\beta_m \geq 0$ and $\sum_m \beta_m = 1$

- naive approach: $\mathbf{K}^* = \frac{1}{M} \sum_m \mathbf{K}^m$

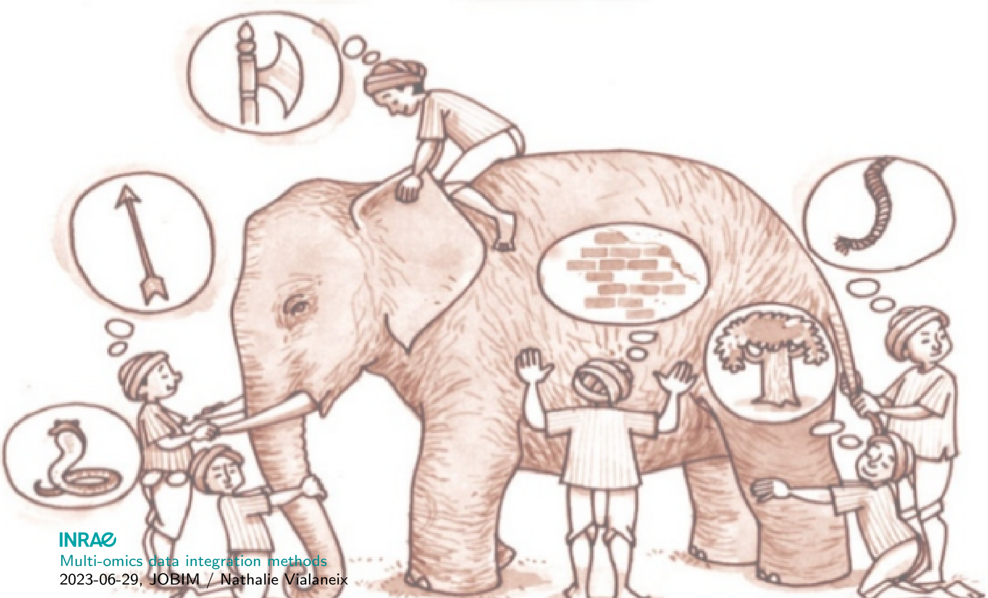
Multiple kernel (or distance) integration

How to "optimally" combine several kernel datasets?

For kernels $\mathbf{K}^1, \ldots, \mathbf{K}^M$ obtained on the same n objects, search: $\mathbf{K}_\beta = \sum_{m=1}^M \beta_m \mathbf{K}^m$ with $\beta_m \geq 0$ and $\sum_m \beta_m = 1$

- naive approach: $\mathbf{K}^* = \frac{1}{M} \sum_m \mathbf{K}^m$
- supervised framework: $\mathbf{K}^* = \sum_m \beta_m \mathbf{K}^m$ with $\beta_m \geq 0$ and $\sum_m \beta_m = 1$ with β_m chosen so as to minimize the prediction error **[Gönen and Alpaydin, 2011]**

Combining kernels in an unsupervised setting



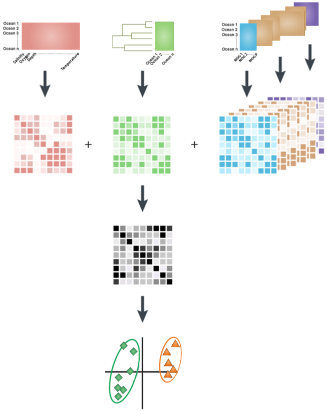
Multiple kernel integration

Ideas of kernel consensus: find a kernel that performs a consensus of all kernels

[Mariette and Villa-Vialaneix, 2018] - R package **mixKernel** with consensus based on:

- STATIS **[L'Hermier des Plantes, 1976, Lavit et al., 1994]**
- criterion that preserves local geometry

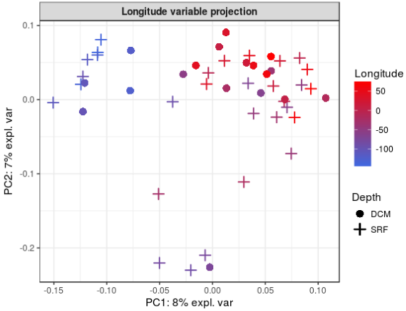
Integrating *TARA* Oceans datasets

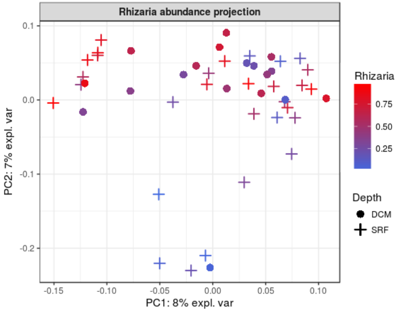


- For all compositional datasets, include phylogenetic information (rather than CLR and alike): weighted Unifrac distance

- Perform PCA (could have been clustering, linear model, . . .) in the feature space.
  + combine with a shuffling approach to identify most influencing variables

Application to *TARA* oceans





Main facts

- Oceans typology related to **longitude**
- *Rhizaria* abundance structure the differences, especially between Arctic Oceans and Pacific Oceans

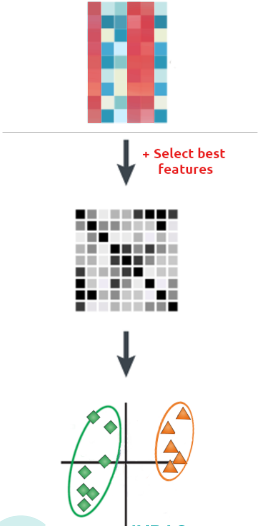
Overview of the talk

Kernel methods

Integrating data with kernels

Conclusion, perspectives

Making kernel methods more interpretable



Which features are important? (for numerical features only)

[Brouard et al., 2022] and **mixKernel**

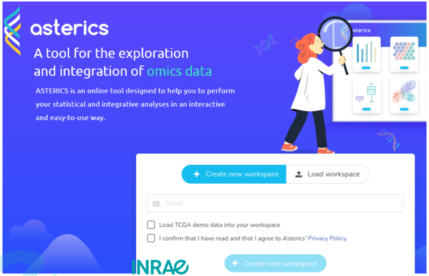
- extension to the unsupervised framework of the work **[Allen, 2013, Grandvalet and Canu, 2002]**
- also extension to the kernel output framework (time series, graph, ... outputs)

Making integration methods available for biologists

Hyphen
from data to science

asterics

http://asterics.miat.inrae.fr



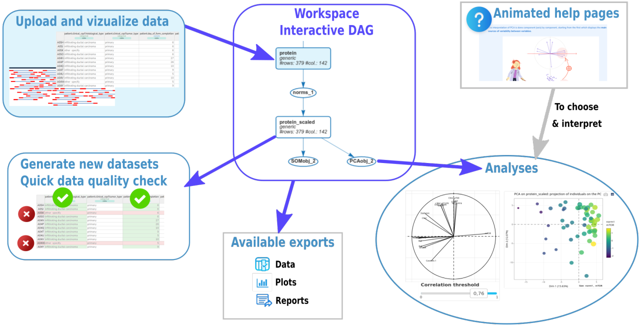
asterics

A tool for the exploration and integration of omics data

ASTERICS is an online tool designed to help you to perform your statistical and integrative analyses in an interactive and easy-to-use way.

□ Load TCGA demo data into your workspace

□ I confirm that I have read and that I agree to *Asterics*' Privacy Policy.



Upload and vizualize data

Workspace Interactive DAG

Animated help pages

To choose & interpret

Generate new datasets
Quick data quality check

Available exports

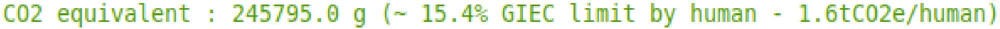
- Data
- Plots
- Reports

Analyses

Future needs for data integration

- improve interpretability of methods (integrate more biological knowledge)

- reduce computational needs to achieve the challenge of a more sustainable research



```
CO2 equivalent : 245795.0 g (~ 15.4% GIEC limit by human - 1.6tCO2e/human)
```

More about machine learning methods?

ECAS

European Courses in Advanced Statistics

SFdS

Random forests: basics, extensions and applications

October 8-13, 2023, Fréjus, France

(with specific classes on random forest for network inference)

Thank you for your attention!

Questions?

References

(unofficial) Beamer template made with the help of Thomas Schiex, Matthias Zytnicki and Andreea Dreau:
https://forgemia.inra.fr/nathalie.villa-vialaneix/bainrae

- Allen, G. I. (2013).
  Automatic feature selection via weighted kernels and regularization.
  Journal of Computational and Graphical Statistics, 22(2):284–299.

- Brouard, C., Mariette, J., Flamary, R., and Vialaneix, N. (2022).
  Feature selection for kernel methods in systems biology.
  NAR Genomics and Bioinformatics, 4(1):lqac014.

- Brouard, C., Shen, H., Dürkop, K., d'Alché Buc, F., Böcker, S., and Rousu, J. (2016).
  Fast metabolite identification with input output kernel regression.
  Bioinformatics, 32(12):i28–i36.

- Foissac, S., Djebali, S., Munyard, K., Vialaneix, N., Rau, A., Muret, K., Esquerre, D., Zytnicki, M., Derrien, T., Bardou, P., Blanc, F., Cabau, C., Crisci, E., Dhorne-Pollet, S., Drouet, F., Faraut, T., Gonzáles, I., Goubil, A., Lacroix-Lamande, S., Laurent, F., Marthey, S., Marti-Marimon, M., Mormal-Leisenring, R., Mompart, F., Quere, P., Robelin, D., SanCristobal, M., Tosser-Klopp, G., Vincent-Naulleau, S., Fabre, S., Pinard-Van der Laan, M.-H., Klopp, C., Tixier-Boichard, M., Acloque, H., Lagarrigue, S., and Giuffra, E. (2019).
  Multi-species annotation of transcriptome and chromatin structure in domesticated animals.
  BMC Biology, 17:108.

- Gönen, M. and Alpaydin, E. (2011).
  Multiple kernel learning algorithms.

Journal of Machine Learning Research, 12:2211–2268.

Grandvalet, Y. and Canu, S. (2002).
Adaptive scaling for feature selection in SVMs.
In Becker, S., Thrun, S., and Obermayer, K., editors, *Proceedings of Advances in Neural Information Processing Systems (NIPS 2002)*, pages 569–576. MIT Press.

Jaakkola, T., Diekhans, M., and Haussler, D. (2000).
A discriminative framework for detecting remote protein homologies.
Journal of Computational Biology, 7(1-2):95–114.

Kondor, R. and Lafferty, J. (2002).
Diffusion kernels on graphs and other discrete structures.
In Sammut, C. and Hoffmann, A., editors, *Proceedings of the 19th International Conference on Machine Learning*, pages 315–322, Sydney, Australia. Morgan Kaufmann Publishers Inc. San Francisco, CA, USA.

Lavit, C., Escoufier, Y., Sabatier, R., and Traissac, P. (1994).
The ACT (STATIS method).
Computational Statistics and Data Analysis, 18(1):97–119.

L'Hermier des Plantes, H. (1976).
Structuration des tableaux à trois indices de la statistique.
PhD thesis, Université de Montpellier.
Thèse de troisième cycle.

Mariette, J. and Villa-Vialaneix, N. (2018).
Unsupervised multiple kernel learning for heterogeneous data integration.
Bioinformatics, 34(6):1009–1015.

Ritchie, M. D., Holzinger, E. R., Li, R., Pendergrass, S. A., and Kim, D. (2015).

Methods of integrating data to uncover genotype-phenotype interactions.
Nature Reviews Genetics, 16(2):85–97.

Saigo, H., Vert, J.-P., Ueda, N., and Akutsu, T. (2004).
Protein homology detection using string alignment kernels.
Bioinformatics, 20(11):1682–1689.

Shen, H., Dührkop, K., Böcher, S., and Rousu, J. (2014).
Metabolite identification through multiple kernel learning on fragmentation trees.
Bioinformatics, 30(12):i157–i64.