Cartography of biostatistics and machine learning methods to identify prognostic factors

Lisa Mounier, Alexandre Civet, Julien Dupin, David Pau, Cyril Esnault

Roche, France

BACKGROUND AND OBJECTIVE

The detection of prognostic factors is a real challenge in healthcare to impact patients journeys for which rigor methodologies are needed. With the emergence of machine learning (ML), new opportunities arise for prognostic factors identification. Although articles exist to review biostatistical methods for the identification of prognostic factors, the opportunities offered by ML algorithms are poorly considered.

The overall purpose is to gather literature and cartography all methods in these two fields that are applicable to identify prognostic factors.

METHODS

A literature review based on relevant keywords has been performed on Google Scholar and PubMed, such as “prognostic factors”, “identification prognostic factors”, ”variables selection” and ”feature selection” combined with keywords like “methods” or “algorithms”. The criteria used for selecting methodological papers also included the date of publication and number of citations. An iterative selection process was then conducted to make an in-depth search and identify new keywords, leading to more specific papers.

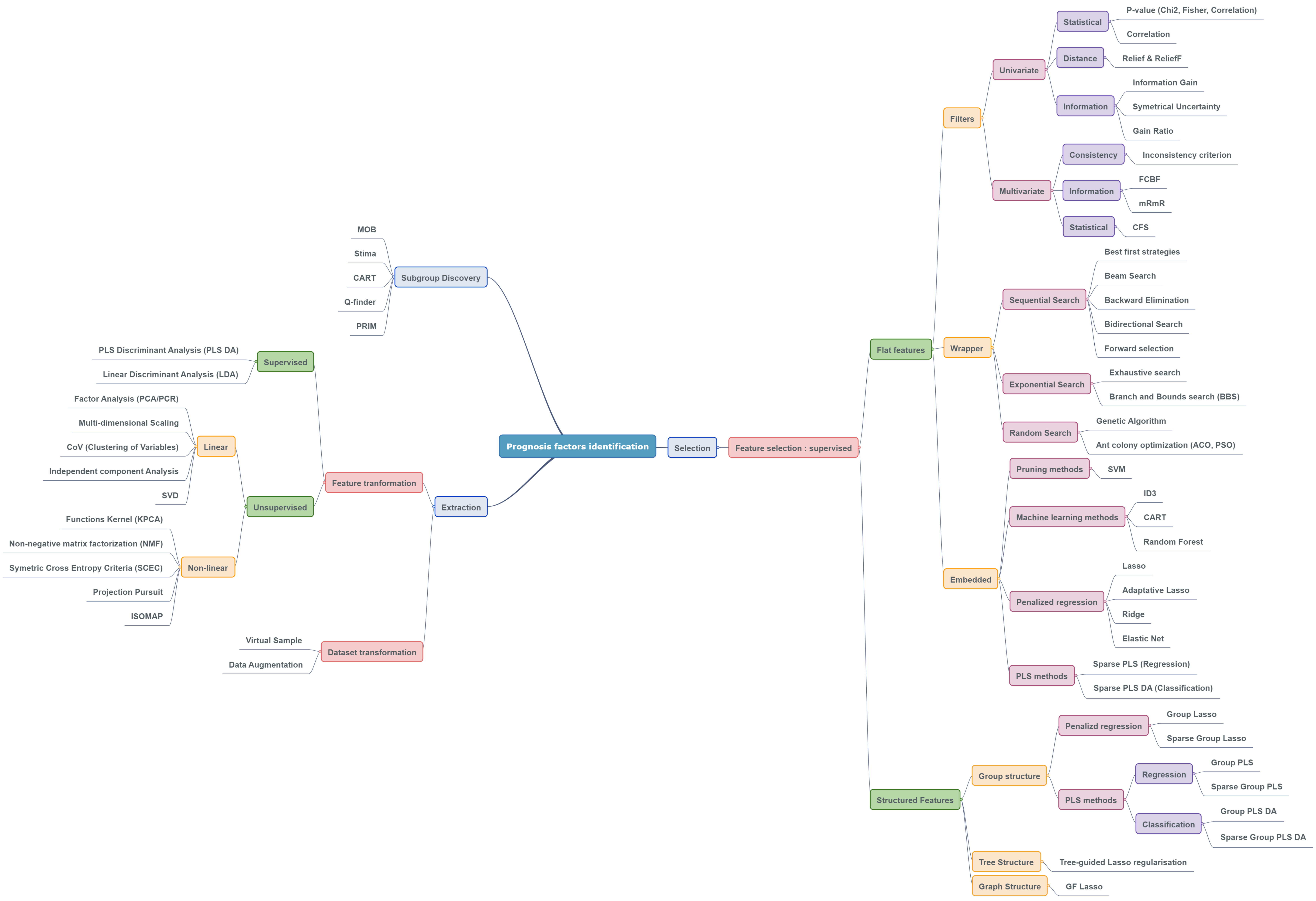
RESULTS

The literature review allowed to identify 25 articles, and 15 were retained for analysis. Based on this synthesis of the literature review, we created a mindmap that covers the 3 steps of prognostic factors identification : feature extraction, feature selection and subgroup discovery fields. Feature selection methods include 3 families for independent features:

1. **Filter** methods that includes univariate and multivariate analysis, which select features based on a criteria (e.g. p-value).
2. **Wrapper** methods which combines iterative search for optimal subset with evaluation through an algorithm (e.g. classifier). These methods include sequential search (e.g. stepwise methods), random search (e.g. genetic algorithms) and exponential search (exhaustive methods) which requires a lot of resources.
3. **Embedded** methods which corresponds to methods where the variable selection process is inherent to the algorithm itself (e.g. lasso regression, random forests). The feature selection is made in the training phase of the model, such as for the selection on node splits for random forests and for setting to 0 the least relevant coefficients for the Lasso regression.

Hybridization of these methods can also be implemented, mixing wrapper and filter methods. Dedicated methods exist for structured features, which allow to integrate knowledge in models on the structure of features in order to identify the most important factors.

Feature extraction includes methods that transform variables or dataset and must be associated with interpretative methods for prognostic factors identification purposes. Finally, subgroup discovery methods include exploratory data mining techniques to uncover patterns associated with an outcome.



Mindmap of methods for prognostic factors identification, covering feature extraction, feature selection and subgroup discovery methods

Conclusion

This research gives an overview of most of existing approaches to identify prognostic factors, both in biostatistics and ML, and highlights that there is a great diversity of approaches. This cartography should help data experts to identify relevant methods and to go beyond what is usually made in studies.

REFERENCE

1. Rodriguez-Galiano VF, Luque-Espinar JA, Chica-Olmo M, Mendes MP. Feature selection approaches for predictive modelling of groundwater nitrate pollution: An evaluation of filters, embedded and wrapper methods. Sci Total Environ. 2018 May 15;624:661-672. doi: 10.1016/j.scitotenv.2017.12.152. Epub 2017 Dec 27. PMID: 29272835.
2. A. Jović, K. Brkić and N. Bogunović, "A review of feature selection methods with applications," 2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2015, pp. 1200-1205, doi: 10.1109/MIPRO.2015.7160458.
3. Tang, J., Alelyani, S., & Liu, H. (2014). Feature selection for classification: A review. In *Data Classification: Algorithms and Applications* (pp. 37-64). CRC Press. https://doi.org/10.1201/b17320
4. Wang L, Wang Y, Chang Q. Feature selection methods for big data bioinformatics: A survey from the search perspective. Methods. 2016 Dec 1;111:21-31. doi: 10.1016/j.ymeth.2016.08.014. Epub 2016 Aug 31. PMID: 27592382.
5. Esra'a Alhenawi, Rizik Al-Sayyed, Amjad Hudaib, Seyedali Mirjalili,Feature selection methods on gene expression microarray data for cancer classification: A systematic review,Computers in Biology and Medicine,Volume 140,2022,105051,ISSN 0010-4825, https://doi.org/10.1016/j.compbiomed.2021.105051.
6. Jia, W., Sun, M., Lian, J., & Hou, S. (2022). Feature dimensionality reduction: a review. *Complex & Intelligent Systems*, 1-31.
7. Halabi S, Owzar K. The importance of identifying and validating prognostic factors in oncology. Semin Oncol. 2010 Apr;37(2):e9-18. doi: 10.1053/j.seminoncol.2010.04.001. PMID: 20494694; PMCID: PMC2929829.
8. Cyril Esnault, May-Line Gadonna, Maxence Queyrel, Alexandre Templier, Jean-Daniel Zucker. Q-Finder: An Algorithm for Credible Subgroup Discovery in Clinical Data Analysis An Application to the International Diabetes Management Practice Study (IDMPS). *Frontiers in Artificial Intelligence and Applications*, IOS Press, 2020, 3, (10.3389/frai.2020.559927). (hal-03155476)