Ensemble methods and online learning for creation and update of prognostic scores in HF patients

Benoît Lalloué, Jean-Marie Monnez

Webinar FIGHT-HF; November 30, 2020

Summary

- How to create a parsimonious event risk score with ensemble methods?

- How to update an ensemble score in the case of a data stream?
  - Tools for generalized linear regression: stochastic approximation processes.

Parsimonious scores by an ensemble method
Context

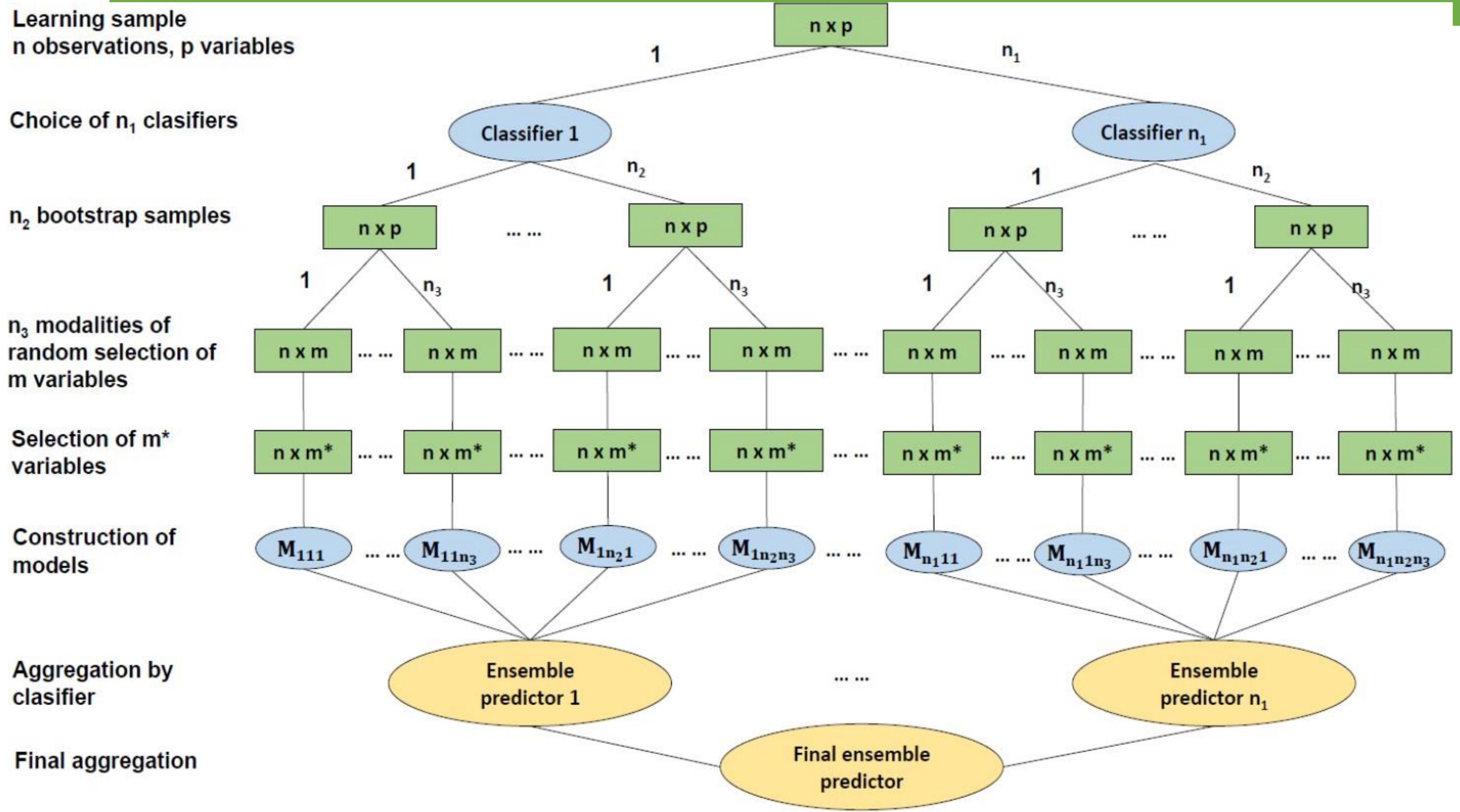
- Scores are mainly built using “classic” statistical methods : logistic regression, Cox regression…

- Another possibility: use ensemble methods.

- Ensemble method: collection of predictors (with different learning rules, samples, selection of variables, etc.) whose predictions are then aggregated.

- Often obtain better results than individual predictors.

Parsimonious scores by an ensemble method
Batch method – Duarte *et al.* 2018



Duarte K, Monnez JM, Albuisson E. Methodology for Constructing a Short Term Event Risk Score in Heart Failure Patients. *Appl Math*. 2018.

Parsimonious scores by an ensemble method
Context (2)

- Common difficulty in the construction of prognostic scores: choose the variables to include.

- Balance between better statistical fit and practical application.

- As we want to use an ensemble method, usual selection methods are not easily applicable.

→ Methodology for constructing parsimonious event scores combining a stepwise preselection of variables and the use of ensemble scores

Parsimonious scores by an ensemble method
Selection methods

- We proposed several methods and compared them.

- Backward methods (need a score formula):
  - Build an ensemble score with a large number of variables
  - Backward selection of the variables, based on the coefficients in the score

- Forward methods (do not need a score formula):
  - Forward selection of the variables which maximize AUC

- A preselection of variables by classifier can precede the methods

Parsimonious scores by an ensemble method
Illustration for short-term predictions in chronic HF patients

- Data: subsample of the GISSI-HF trial
- Data management: couples patient-visit; winsorized and transformed variables; balancing of the sample (duplication of the cases)
- Event: hospitalization for aggravating HF or death from HF within 180 days of a visit
- 3 methods compared: similar selections of variables and performances
- 4 parsimonious scores using the fastest method:

Score's name	S3.26	S3.15	S3.8	S3.2
Nb of variables used	26	15	8	2
AUC OOB final score	0.8137	0.8002	0.7835	0.7523

Online logistic regression
Online learning & online standardization

Online learning:

- Analysis of a data stream or of big data.
- Update the results in successive steps, taking into account new data at each step.
- A possibility: use recursive stochastic algorithms.

Online standardization of the data:

- Data can be standardized to: avoid a numerical explosion or apply a shrinkage method (e.g. LASSO).
- Issue for data streams: means and variances are a priori unknown.
- A possibility: do an online standardization.
- Studied for the linear regression: better performance compared to raw data.
- We used a similar approach for the logistic regression.

Online logistic regression
Stochastic gradient processes

Stochastic approximation processes of this form were tested:

$$X_{n+1} = X_n - a_n \frac{1}{m_n} \sum_{j \in I_n} \tilde{Z}_j \left(h\left(\tilde{Z}_j{}' X_n \right) - S_j \right)$$

$$\bar{X}_{n+1} = \frac{1}{n+1} \sum_{i=1}^{n+1} X_i$$

Different variants exist:

- Classical (X_n) or averaged $(\bar{X}_n)$.
- Raw data or online standardized data.
- Different numbers of new observations at each step (m_n).
- Variable step-size or piecewise constant step-size (a_n).

Online logistic regression
Datasets, datastream & comparison

- 24 processes tested on 5 datasets. Data streams simulated by randomly drawing successive data batches from the datasets.

- Usual logistic regression used as gold standard.

- Convergence criterion (norms ratio: $\frac{\|\theta^c - \widehat{\theta}_{n+1}\|}{\|\theta^c\|}$) recorded for fixed numbers of observations used and for fixed processing times.

- Processes ranked for each dataset and each recording point. Average rank across all datasets used to compare processes.

Online logistic regression
Comparison for a fixed processing time (60s)

Process	Twonorm	Ringnorm	Quantum	Adult	HOSPHF30D	Mean rank
CR1V	0.055	0.019*	0.288	EXPL	EXPL	-
CR10V	0.061	0.005*	0.310	EXPL	EXPL	-
CR100V	0.073	0.002*	0.333	EXPL	EXPL	-
AR1P50	0.011*	0.019*	0.086	EXPL	EXPL	-
AR10P50	0.002*	0.002*	0.095	EXPL	EXPL	-
AR100P50	0.001*	0.001*	0.102	EXPL	EXPL	-
AR1P100	0.015*	0.029*	0.064	EXPL	EXPL	-
AR10P100	0.002*	0.003*	0.079	EXPL	EXPL	-
AR100P100	0.001*	0.001*	0.090	EXPL	EXPL	-
AR1P200	0.018*	0.052	0.040*	EXPL	EXPL	-
AR10P200	0.002*	0.005*	0.064	EXPL	EXPL	-
AR100P200	0.001*	0.001*	0.076	EXPL	EXPL	-
CS1V	0.139	0.023*	0.173	0.134	0.153	10.0
CS10V	0.182	0.011*	0.057	0.101	0.228	9.0
CS100V	0.227	0.004*	0.071	0.108	0.326	9.0
AS1P50	0.027*	0.025*	0.042*	0.389	0.095	8.6
AS10P50	0.006*	0.005*	0.014*	0.020*	0.053	4.8
AS100P50	0.009*	0.002*	0.007*	0.017*	0.014*	3.2
AS1P100	0.032*	0.037*	0.071	0.386	0.087	9.2
AS10P100	0.005*	0.006*	0.014*	0.025*	0.050*	4.8
AS100P100	0.004*	0.002*	0.007*	0.011*	0.011*	1.8
AS1P200	0.046*	0.060	0.121	0.498	0.112	10.6
AS10P200	0.005*	0.008*	0.017*	0.035*	0.049*	5.4
AS100P200	0.003*	0.002*	0.007*	0.009*	0.012*	1.6

* Denotes a criterion value <0.05

EXPL: numerical explosion

- *Process type:* C for classical SGD, A for ASGD
- *Data type:* R for raw, S for online standardized
- *1st number:* number of new obs. per step
- *Step-size:* V for variable, P for piecewise constant (*2nd number:* levels size)

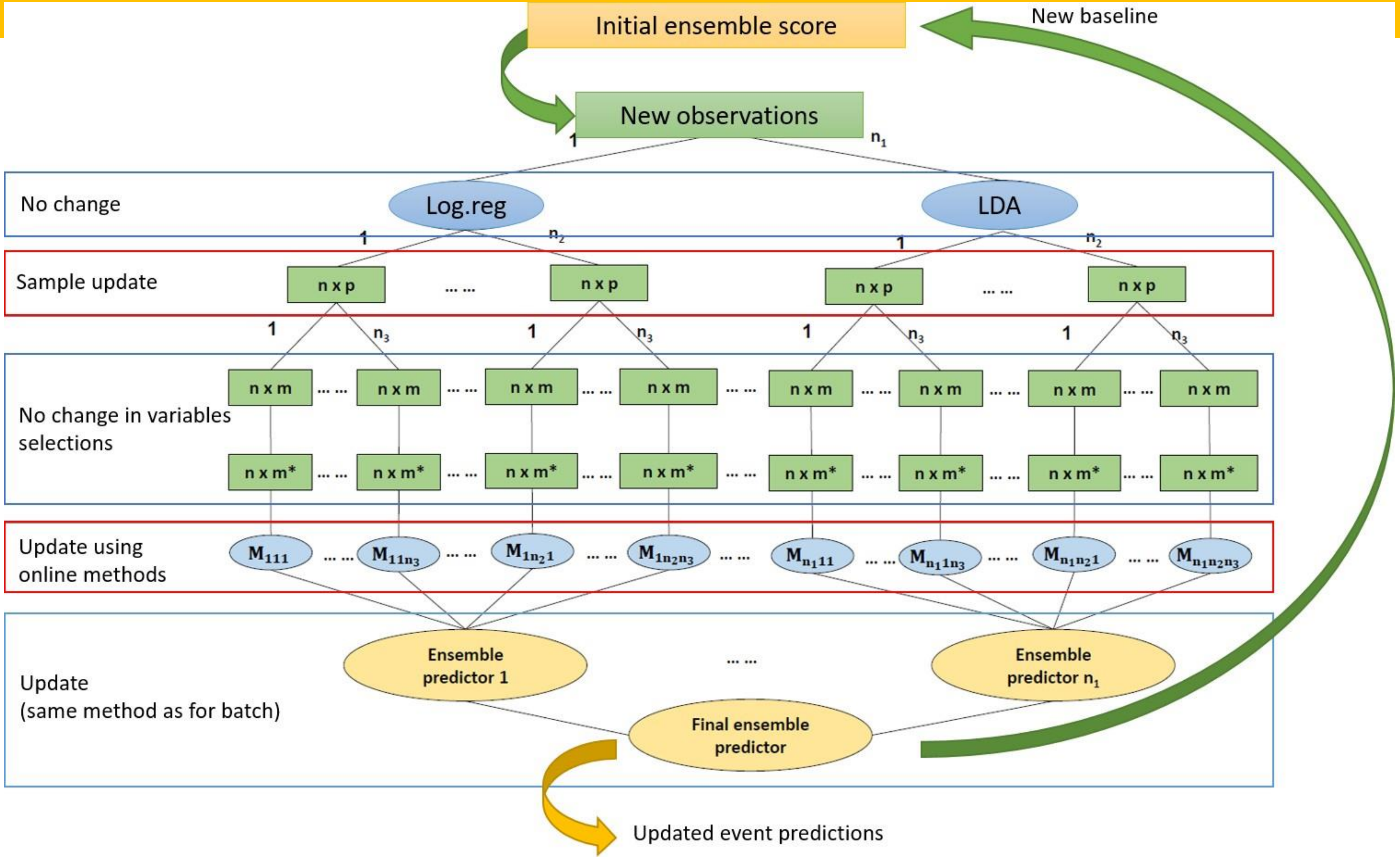
Online ensemble score
Online method

How to update an ensemble score similar to Duarte et al. in the case of a data stream?

- Choice of classifiers: same as the initial ensemble score.
- Bootstrap samples: use Poisson bootstrap.
- Selection of variables: same as the initial ensemble score.
- Construction of models: use online versions (online linear regression, online logistic regression...).
- Aggregation: same as the initial ensemble score.

Online ensemble score

Online method (2)



Online ensemble score

Experiments

- Same datasets than previously. Data streams simulated by randomly drawing successive data batches from the datasets.

- A batch score was created as reference for each dataset:
  - 100 bootstrap samples.
  - 2 classifiers: logistic regression and linear discriminant analysis (linear regression).
  - 1 modality with all variables.

- 6 online scores using 100N observations and the same parameters.

- Empirical study of convergence toward the reference score ($\frac{\|\theta^c - \widehat{\theta}_{n+1}\|}{\|\theta^c\|}$).

Online ensemble score

Comparison with a fixed number of observations (100N)

Norms ratio between the batch score coefficients and the online scores coefficients:

Process		Twonorm	Ringnorm	Quantum	Adult	HOSPHF30D
CS100V_CS100V	*LDA*	0.0010*	0.0020*	0.0073*	0.0076*	0.0165*
	Log. Reg.	0.0033*	0.0009*	0.0168*	0.1002	0.0566
	Final	**0.0015***	**0.0014***	**0.0083***	**0.0414***	**0.0289***
AS100P50_AS100P50	*LDA*	0.0006*	0.0007*	0.0027*	2.7560	0.0176*
	Log. Reg.	0.0006*	0.0007*	0.0032*	0.0346*	0.0203*
	Final	**0.0005***	**0.0007***	**0.0029***	**1.6968**	**0.0192***
AS100C_AS100P200	*LDA*	0.0006*	0.0007*	0.0028*	0.0066*	0.0165*
	Log. Reg.	0.0007*	0.0007*	0.0033*	0.0069*	0.0206*
	Final	**0.0006***	**0.0007***	**0.0030***	**0.0067***	**0.0190***
CS100Vall_CS100V	*LDA*	0.0005*	0.0006*	0.0033*	0.0287*	0.0153*
	Log. Reg.	0.0033*	0.0009*	0.0168*	0.1002	0.0566
	Final	**0.0017***	**0.0007***	**0.0090***	**0.0281***	**0.0290***
AS100P50all_AS100P50	*LDA*	0.0006*	0.0007*	0.0046*	0.0100*	0.0060*
	Log. Reg.	0.0006*	0.0007*	0.0032*	0.0346*	0.0203*
	Final	**0.0005***	**0.0007***	**0.0039***	**0.0193***	**0.0147***
AS100Call_AS100P200	*LDA*	0.0006*	0.0007*	0.0046*	0.0153*	0.0060*
	Log. Reg.	0.0007*	0.0007*	0.0033*	0.0069*	0.0206*
	Final	**0.0005***	**0.0007***	**0.0039***	**0.0120***	**0.0149***

Conclusion

Parsimonious scores:

- Methods which build a succession of scores from which the user can choose according to its objectives.
- In the application: similar or better results than other scores, with less variables.

Online logistic regression:

- Online standardization of the data helps to avoid numerical explosion.
- Interest of averaged processes with piecewise constant step-size and online standardized data.

Online ensemble score:

- Online ensemble scores converge empirically to the batch score (theoretical convergence already proven).

Conclusion

References

Parsimonious ensemble score:

Lalloué B, Monnez JM. Construction of parsimonious event risk scores by an ensemble method. An illustration for short-term predictions in chronic heart failure patients. 2020. (in preparation, submission planned in PLOS One)

Online logistic regression:

Lalloué B, Monnez JM, Albuisson E. Régression logistique sous contrainte avec standardisation en ligne pour flux de données. 26èmes Rencontres de la SFC. 2019.

Lalloué B, Monnez JM, Albuisson E. Streaming constrained binary logistic regression with online standardized data. 2020. hal-02156324 (minor revision in *Journal of Applied Statistics*)

Online score:

Lalloué B, Monnez JM, Albuisson E. Actualisation en ligne d’un score d’ensemble. 51e Journées de Statistique. 2019.

Lalloué B, Monnez JM, Albuisson E. Convergence d'un score d'ensemble en ligne : étude empirique. 52e Journées de Statistique. 2020.

Lalloué B, Monnez JM. Construction and update of an online ensemble score. 2020 (in preparation)