S2. Additional figures and table

2021-10-21

Contents

Posterior predictive distributions

Figure 1 show the mixture of the PPD of the probability of a positive match, grouped by label (only records manually labelled or reviewed are considered). The posterior samples for each record were extracted and joined into a global distribution; on this distribution, density was computed on the logit scale and then logistic transformed for displaying on a $[0-1]$ scale. The purple ridges show the distribution of the still unlabelled records.
For each iteration, the thresholds which define the "uncertainty zone", i.e, the lower and the upper range of 98% PrI for the positive and negative records respectively, are shown. Records whose 98% PrI intersects the uncertainty zone requires manual review.
Notice how the positive and negative record densities tend to increasingly overlap at each iteration; meanwhile the distribution of the records to be reviewed shrinks and shifts towards the negative side, as positive records get found and labeled.
In Session 2, as the number of negative records reviewed increased, also already positively labelled records re-entered the uncertainty zone; this is due to the baseline positivity rate decreasing as the number of negatives in the training data increased, forcing to review dubious records that may have been mislabelled.

List of terms relevant for prediction

In table 1 and 2 are listed the 50 more relevant terms used by the BART algorithm to discriminate between positive and negative records, for Session 1 and 2 (see Methods). Term importance ("Inclusion rate" in the tables) is defined as the ensemble average inclusion rate in posterior trees over 10,000 total term inclusions, while the Inclusion Stability (IS) is the ratio of the average inclusion rate over its standard deviation among the ensemble models. The symbol "|" in the terms indicate redundant terms while "&" indicate nc-grams. The component in which the term was used is reported in the leftmost column.
For each term, we added its linear association with a positive label estimated through a Poisson regression, reporting it as Relative Risk (RR) and its statistical significance (Statistic) measured as number of standard errors, s.e., of the terms. A strong BART score with a Statistic close to zero identify terms whose effect is highly non-linear (e.g., they are relevant only in the presence of other terms).

Hyperparameters grid search

To select the hyperparameter set which would maximize sensitivity and efficiency during classification, we set up a comprehensive grid search by running the algorithm on a fully labeled validation dataset of 1200 records. Using a partition tree algorithm we grouped the searches in a number of clusters with similar performance and selected the best parameter set in the best cluster.
Table 3 shows the best hyperparameter set and their performance for each cluster while figure 2 displays the conditional impact on performance of each hyperparameter

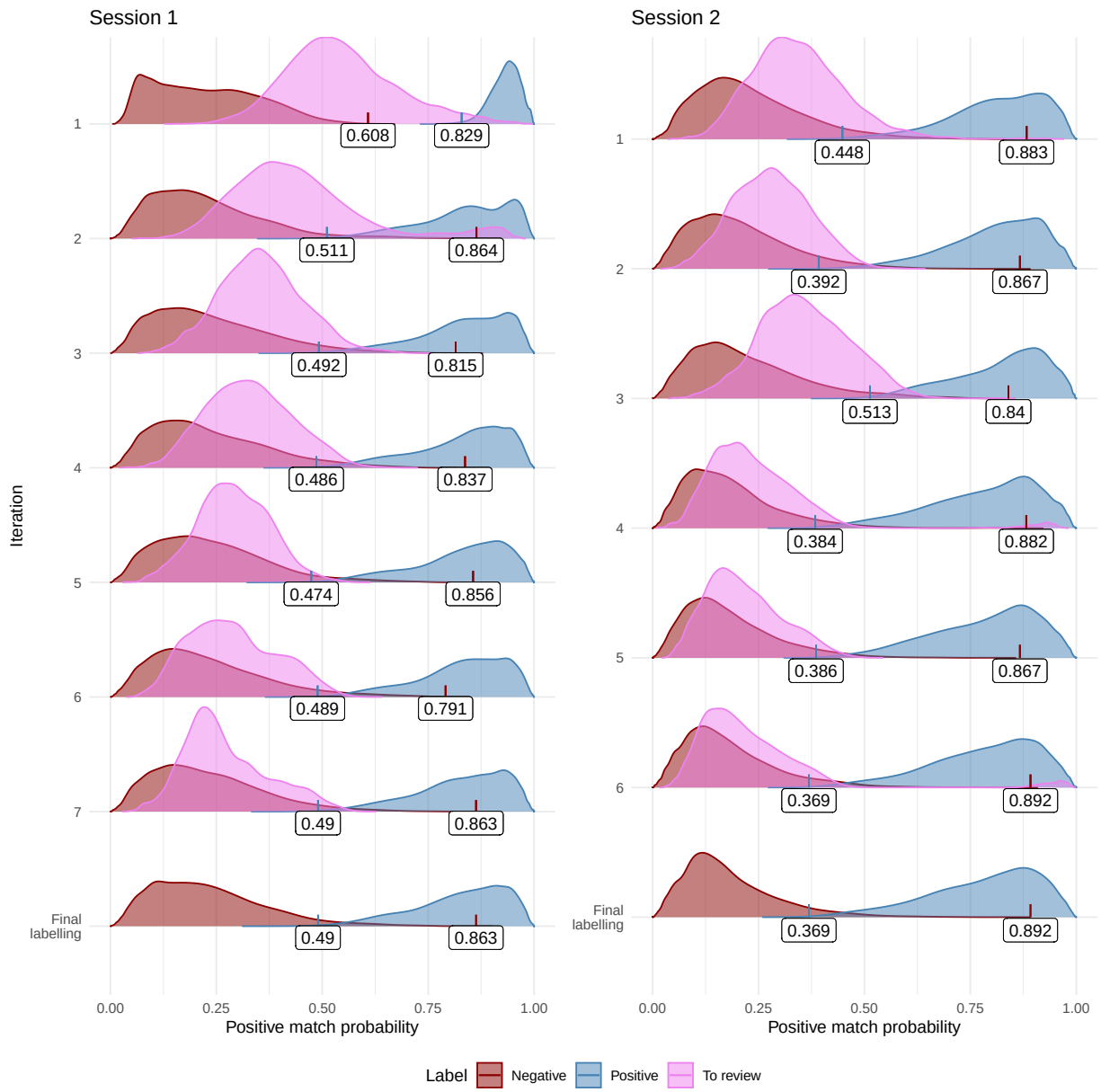

Figure 1. Mixture predictive distribution of the probability of a positive match, grouped by labelling status.

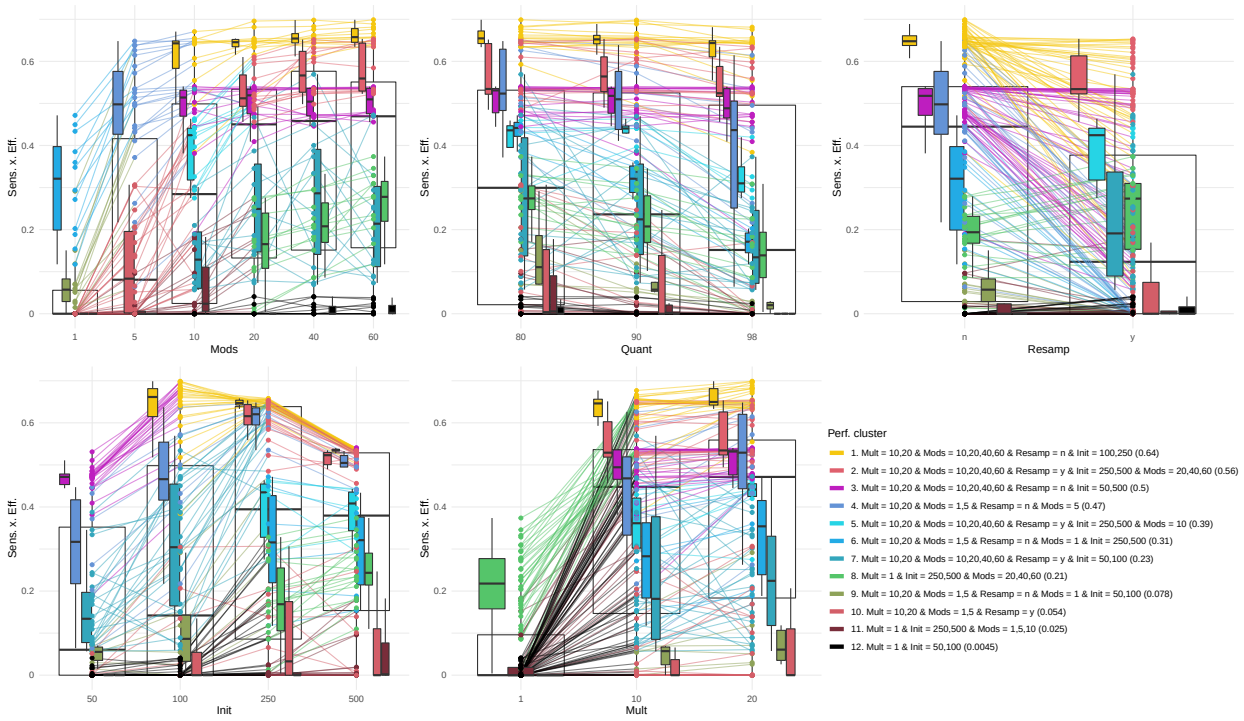

Figure 2. Performance clusters and impact of single hyperparameters on engine performance. The performance is measured as Sensitivity x Efficiency. Each cluster is color coded. Mods: num. of models in the ensemble; Quant: PrI limits for the uncertainty zone; Resamp: bootstrap resampling; Init: num. of initial training set; Multi: oversampling multiplier of positive matches.

Table 1. Term importance at the end of Session 1.

Component	Term	Inclusion rate	IS	RR	Statistic
Keyword	Patient Transport	61.2	3.77	99.1	21.3
Abstract	Transfer	57.0	3.93	22.5	15.4
Title	Network	56.5	2.91	18	14.2
Abstract	Network & Patient	54.2	4.66	26.3	15.2
Author	Donker T	53.5	4.56	159	16.5
Abstract	Worker	50.0	3.33	0.421	-1.21
Keyword	Hospitals	49.8	4.31	27.8	16.5
Abstract	Movement	47.8	2.70	27.2	15
Title	Spread	46.6	2.25	16.2	12.1
Abstract	Facility	45.0	2.22	19.6	14.8
Keyword	Orange County	44.3	3.19	199	17.2
Abstract	Conduct	42.6	3.70	0.221	-2.57
Abstract	Patient	42.0	3.61	27.6	7.23
Abstract	Perform	41.9	2.38	0.342	-2.55
Title	Hospital	39.0	1.95	12.5	12.5
Abstract	Regional	38.9	3.08	21.7	14.9
Abstract	Agent	38.1	2.74	4.36	6.28
Abstract	California	37.3	2.60	38	12.6
Title	Transfer	36.6	3.54	27	11.8
Keyword	Patient Transfer | Patient Transfer & Patient Transport	36.6	2.16	164	2
Abstract	Finding	33.2	1.78	0.372	-2.35
Title	Outbreak	32.9	2.91	3.4	3.51
Abstract	Collect	32.4	1.82	0.408	-1.95
Title	Regional	32.3	2.68	44.2	14.2
Abstract	Network	31.6	3.09	12.8	11.7
Abstract	Resistant	31.6	2.55	11	11.2
Abstract	Outcome	31.2	1.93	0.178	-2.95
Abstract	Discharge	31.1	3.79	9.99	9.02
Abstract	2014	30.3	1.84	0.588	-1.04
Abstract	Practice	29.5	2.47	0.508	-1.33
Abstract	Culture	28.9	2.14	0.378	-1.66
Abstract	Positive	28.8	1.96	0.346	-2.52
Abstract	Gene	28.3	1.29	2.4e-07	-0.0415
Abstract	Disease	28.0	2.42	0.365	-3.7
Keyword	Enterococci	27.5	1.92	25.2	6.33
Abstract	Month	27.3	1.86	0.288	-2.44
Abstract	Healthcare & Facility	26.2	1.74	34.9	17
Abstract	Prevalence	26.2	1.71	4.08	6.69
Abstract	Effort	26.1	1.22	6.64	8.57
Abstract	Length	25.3	1.44	6.1	6.45
Keyword	System	25.0	1.43	11.9	3.47
Abstract	Laboratory	24.4	1.38	0.442	-1.39
Keyword	Resistant Staphylococcus Aureus	24.4	1.59	22.8	11.2
Abstract	Clinical	23.6	2.10	0.421	-2.9
Abstract	Dataset	22.5	1.47	8.75	6.5
Abstract	Development	22.2	1.22	0.148	-2.68
Abstract	Hand	22.0	1.97	0.822	-0.335
Keyword	Pathogen Transmission	21.8	3.26	67.9	7.2
Keyword	Cross Infection & Humans & Transmission	21.7	1.40	31.1	15.5
Abstract	Flow	21.4	1.14	4.07	4.56

Table 2. Term importance at the end of Session 2.

Component	Term	Inclusion rate	IS	RR	Statistic
Keyword	Patient Transport	61.2	3.77	99.1	21.3
Abstract	Transfer	57.0	3.93	22.5	15.4
Title	Network	56.5	2.91	18	14.2
Abstract	Network & Patient	54.2	4.66	26.3	15.2
Author	Donker T	53.5	4.56	159	16.5
Abstract	Worker	50.0	3.33	0.421	-1.21
Keyword	Hospitals	49.8	4.31	27.8	16.5
Abstract	Movement	47.8	2.70	27.2	15
Title	Spread	46.6	2.25	16.2	12.1
Abstract	Facility	45.0	2.22	19.6	14.8
Keyword	Orange County	44.3	3.19	199	17.2
Abstract	Conduct	42.6	3.70	0.221	-2.57
Abstract	Patient	42.0	3.61	27.6	7.23
Abstract	Perform	41.9	2.38	0.342	-2.55
Title	Hospital	39.0	1.95	12.5	12.5
Abstract	Regional	38.9	3.08	21.7	14.9
Abstract	Agent	38.1	2.74	4.36	6.28
Abstract	California	37.3	2.60	38	12.6
Title	Transfer	36.6	3.54	27	11.8
Keyword	Patient Transfer | Patient Transfer & Patient Transport	36.6	2.16	164	2
Abstract	Finding	33.2	1.78	0.372	-2.35
Title	Outbreak	32.9	2.91	3.4	3.51
Abstract	Collect	32.4	1.82	0.408	-1.95
Title	Regional	32.3	2.68	44.2	14.2
Abstract	Network	31.6	3.09	12.8	11.7
Abstract	Resistant	31.6	2.55	11	11.2
Abstract	Outcome	31.2	1.93	0.178	-2.95
Abstract	Discharge	31.1	3.79	9.99	9.02
Abstract	2014	30.3	1.84	0.588	-1.04
Abstract	Practice	29.5	2.47	0.508	-1.33
Abstract	Culture	28.9	2.14	0.378	-1.66
Abstract	Positive	28.8	1.96	0.346	-2.52
Abstract	Gene	28.3	1.29	2.4e-07	-0.0415
Abstract	Disease	28.0	2.42	0.365	-3.7
Keyword	Enterococci	27.5	1.92	25.2	6.33
Abstract	Month	27.3	1.86	0.288	-2.44
Abstract	Healthcare & Facility	26.2	1.74	34.9	17
Abstract	Prevalence	26.2	1.71	4.08	6.69
Abstract	Effort	26.1	1.22	6.64	8.57
Abstract	Length	25.3	1.44	6.1	6.45
Keyword	System	25.0	1.43	11.9	3.47
Abstract	Laboratory	24.4	1.38	0.442	-1.39
Keyword	Resistant Staphylococcus Aureus	24.4	1.59	22.8	11.2
Abstract	Clinical	23.6	2.10	0.421	-2.9
Abstract	Dataset	22.5	1.47	8.75	6.5
Abstract	Development	22.2	1.22	0.148	-2.68
Abstract	Hand	22.0	1.97	0.822	-0.335
Keyword	Pathogen Transmission	21.8	3.26	67.9	7.2
Keyword	Cross Infection & Humans & Transmission	21.7	1.40	31.1	15.5
Abstract	Flow	21.4	1.14	4.07	4.56

Table 3. Hyperparameter clusters and best cluster subsets. For each cluster, the defining rules and mean Sens. x Eff. is shown, followed by the per-cluster best set results.

Cluster (mean score)	Num. iterations	Positive matches	Reviewed records	Sensitivity	Efficiency	Score (Sens. x Eff.)	Num. ensemble models	Uncertainty interval	Resampling	Num. initial training records	Positives oversampling multiplier
1. Mult = 10,20 & Mods = 10,20,40,60 & Resamp = n & Init = 100,250 (0.64)	7	81 / 82	462 / 1200	98.8%	61.5%	0.608	10	98	n	250	10
2. Mult = 10,20 & Mods = 10,20,40,60 & Resamp = y & Init = 250,500 & Mods = 20,40,60 (0.56)	5	82 / 82	618 / 1200	100%	48.5%	0.485	20	80	y	250	10
3. Mult = 10,20 & Mods = 10,20,40,60 & Resamp = n & Init = 50,500 (0.5)	6	81 / 82	589 / 1200	98.8%	50.9%	0.503	10	98	n	500	10
4. Mult = 10,20 & Mods = 1,5 & Resamp = n & Mods = 5 (0.47)	8	82 / 82	1123 / 1200	100%	6.42%	0.0642	5	98	n	50	10
5. Mult = 10,20 & Mods = 10,20,40,60 & Resamp = y & Init = 250,500 & Mods = 10 (0.39)	5	82 / 82	685 / 1200	100%	42.9%	0.429	10	80	y	250	10
6. Mult = 10,20 & Mods = 1,5 & Resamp = n & Mods = 1 & Init = 250,500 (0.31)	5	82 / 82	874 / 1200	100%	27.2%	0.272	1	90	n	250	10
7. Mult = 10,20 & Mods = 10,20,40,60 & Resamp = y & Init = 50,100 (0.23)	7	82 / 82	948 / 1200	100%	21%	0.21	10	90	y	50	20
8. Mult = 1 & Init = 250,500 & Mods = 20,40,60 (0.21)	5	82 / 82	806 / 1200	100%	32.8%	0.328	60	80	y	250	1
9. Mult = 10,20 & Mods = 1,5 & Resamp = n & Mods = 1 & Init = 50,100 (0.078)	6	82 / 82	1139 / 1200	100%	5.08%	0.0508	1	90	n	50	20
10. Mult = 10,20 & Mods = 1,5 & Resamp = y (0.054)	5	82 / 82	832 / 1200	100%	30.7%	0.307	5	80	y	250	20
11. Mult = 1 & Init = 250,500 & Mods = 1,5,10 (0.025)	5	82 / 82	981 / 1200	100%	18.2%	0.182	10	90	y	500	1
12. Mult = 1 & Init = 50,100 (0.0045)	6	82 / 82	1151 / 1200	100%	4.08%	0.0408	20	80	y	50	1