A Flood of Information: Leveraging Machine Learning to Augment Screening Decisions in Ever-Expanding Systematic Reviews.

William Letton;[1] Chris Martin[1]
[1] Crystallise Ltd, London, UK william.letton@crystallise.com, chris.martin@crystallise.com

Leave Empty

This space will be automatically filled with a QR code and number for easy sharing

Introduction

- Abstract screening during systematic reviews is error-prone, time-consuming, expensive, and requires expert judgement. These problems are increasing as research data volume increases exponentially.
- Methods to improve speed and accuracy are therefore desirable.
- A variety of machine learning methods have been developed to address this need[1,2].
- Here we present the development of a program that uses simple machine learning to review human screening and suggest the most likely false negatives for review. In this context text classifiers are used as outlier detectors, to identify human screening decisions that do not fit the overall pattern.

Methods

- The program was developed in the Python™ language.
- Two classifier models were tested:
  - Support vector machine (**SVM**) implemented using scikit-learn[3].
  - Naïve **Bayes**.
- Two class structures were tested:
  - Boolean **assignments** (simply "include" or "exclude")
  - Multiple **tags** (e.g. "include_methods", "exclude_population", "exclude_disease").
- For training classifiers, the abstracts were represented as bag-of-word feature vectors.
- For each screener for each project the classifier was trained using the human screening decisions, and was then used to re-classify those same abstracts, producing an inclusion score for each.

Figures 1.1, 1.2, and 1.3

- The four combinations of classifier model and class structure were tested using datasets from three completed systematic reviews. Receiver operating characteristic (ROC) curves show the ability of each method to find the false negatives left by the human screener among the true negatives. This is also represented as area under the curve (AUC) scores.

Figure 2

- The highest scoring method (SMV with Boolean assignments) was then tested on a 'live' project.
  - 100 real review suggestions were made, along with 100 random ones.
  - The net reclassification index (NRI) was used as a measure of the utility of the suggestions.
  - The ratio of the NRI for the real and random suggestions was used as a measure of the utility of the real suggestions relative to random chance.
  - In cases where the NRI for the random suggestions was 0, the utility ratio was taken as the raw NRI for the real suggestions.

Results

Figure 1.1 – ROC performance of different methods in study 1.

	Human screener	
	A	B
Abstracts Screened	7166	7126
Screening includes	428	571
Screening excludes	6738	6555
P (excludes later reconciled as includes)	164	39
N (excludes not later reconciled as includes)	6574	6516

	Human screener	
AUC by classifier method	A	B
Bayes tags	0.55	0.47
Bayes assignments	0.80	0.90
SVM tags	0.82	0.90
SVM assignments	0.84	0.91

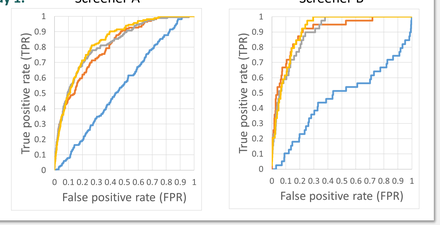

Figure 1 – ROC plots for the four classifier methods over three systematic review projects. AUC values are given in the lower left tables.

ROC plot legend
Bayes tags
Bayes assignments
SVM tags
SVM assignments

Figure 1.2 – ROC performance of different methods in study 2.

	Human screener			
	A	B	C	D
Abstracts Screened	2152	4068	4352	2240
Screening includes	185	243	236	154
Screening excludes	1967	3825	4116	2086
P (excludes later reconciled as includes)	54	17	5	89
N (excludes not later reconciled as includes)	1913	3808	4111	1997

	Human screener			
AUC by classifier method	A	B	C	D
Bayes tags	0.72	0.46	0.52	0.70
Bayes assignments	0.90	0.82	0.81	0.83
SVM tags	0.93	0.72	0.78	0.71
SVM assignments	0.94	0.87	0.92	0.83

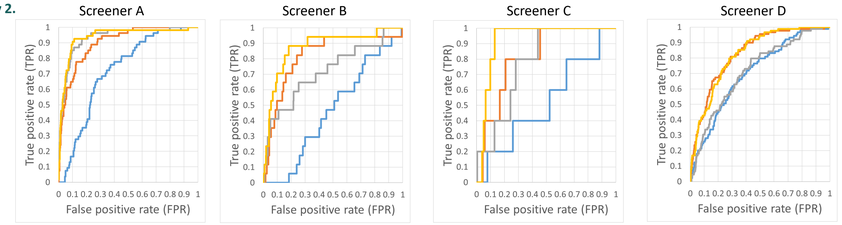

Figure 1.3 – ROC performance of different methods in study 3.

	Human screener			
	A	B	C	D
Abstracts Screened	457	4599	2798	1880
Screening includes	58	572	317	117
Screening excludes	399	4027	2481	1763
P (excludes later reconciled as includes)	2	46	26	30
N (excludes not later reconciled as includes)	397	3981	2455	1733

	Human screener			
AUC by classifier method	A	B	C	D
Bayes tags	0.49	0.65	0.68	0.43
Bayes assignments	0.37	0.75	0.77	0.65
SVM tags	0.42	0.70	0.75	0.65
SVM assignments	0.40	0.73	0.71	0.67

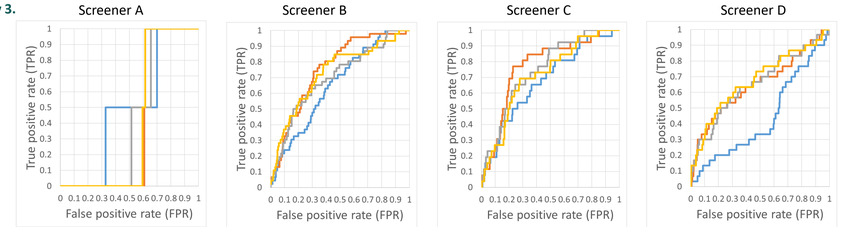

Figure 2– Assessing performance of classifier suggestions vs random suggestions in a live project

	Human screener							
	B	C	D	E	F	G	FINAL	TOTAL
Abstracts Screened	941	1307	852	1198	1586	1557	5027	12468
Screening excludes	844	1197	635	1077	1458	1290	4650	11151
Real suggestions	100	100	100	100	100	100	100	700
Real suggestions reviewed as IN	29	3	0	6	10	24	15	87
Real net reclassification index (%)	29.0%	3.0%	0.0%	6.0%	10.0%	24.0%	15.0%	12.4%
Random suggestions	100	100	100	100	100	100	100	700
Random suggestions reviewed as IN	5	0	0	0	0	3	1	9
Random net reclassification index (%)	5.0%	0.0%	0.0%	0.0%	0.0%	3.0%	1.0%	1.3%
Net reclassification index ratio	5.8	3	1	6	10	8	15	9.7

Figure 2 – Assessing the utility of the SVM assignments classifier in a live project. The net reclassification index (NRI) gives the proportion of suggestions that are useful, while the ratio of the NRIs gives the relative utility of the real suggestions verses ones made at random.

Results

- The results varied between projects and screeners.
- On average the SVM classifier using Boolean assignments performed best, achieving a median discriminator ROC-AUC score of 0.83 across the three test projects.
- Across all six screeners and the 'final' combined screening, 12.4% of the real suggestions resulted in abstract reclassification, while only 1.3% of the random suggestions did.
- Therefore on average a real suggestion was 9.7 times as likely to result in a reclassification as a random one.

Discussion

- Although the utility of this simple 'third screener' program varied with project and human screener, on a live test it produced suggestions that were better than chance in all cases (the only exception being screener D, who made no reclassifications in either case).
- These results demonstrate the utility of machine-learning techniques to augment systematic reviews, by reducing the false-negative rate during screening.
- One major advantage of this approach, compared to other machine-learning approaches, is that all decision-making is still performed by human screeners.
- One major disadvantage of this approach is that it can only detect outliers in otherwise correct screening decisions. If the human screener systematically makes screening errors these will not be detected.
- Further development of this 'third screener' approach could investigate how utility varies with the kind of systematic review project, or between screeners of different levels of expertise in the topic.

References

1. O'Mara-Eves A, Thomas J, McNaught J, Miwa M, Ananiadou S. Using text mining for study identification in systematic reviews: A systematic review of current approaches. Syst Rev. 2015;4(1):5. doi:10.1186/2046-4053-4-5.
2. 1. Marshall IJ, Wallace BC. Toward systematic review automation: A practical guide to using machine learning tools in research synthesis. Syst Rev. 2019;8(1):1-10. doi:10.1186/s13643-019-1074-9.
3. Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

Crystallise Ltd. 19 Saffron Court, Southfields Business Park, Basildon, Essex, UK
Tel: +44 (0)1268 543470
For a copy of this poster, email: william.letton@crystallise.com
www.crystallise.com www.heoro.com

Presented at the SMDM 42nd Annual Virtual Meeting
October 2020;
Entry id #13374

heoro.com