RESEARCH

Statistical Stopping Criteria for Active Learning in Systematic Review Screening

Max W Callaghan[1,2*] and Finn Müller-Hansen[1]

*Correspondence:
callaghan@mcc-berlin.net
[1]Mercator Research Institute on Global Commons and Climate Change, Torgauer Straße, 10829 Berlin, Germany
Full list of author information is available at the end of the article

Abstract

First part title: Active learning for systematic review screening promises to reduce the human effort required to identify relevant documents for a systematic review. Machines and humans work together, with the human providing training data, and the machine optimising the documents a human screens. However, current approaches lack robust stopping criteria, meaning that such systems are hard to implement in live reviews. This paper introduces a workflow for working with robust and flexible statistical stopping criteria, that offer real work reductions on the basis of a given confidence level of reaching a given recall.

Keywords: sample; article; author

Background

Machine learning for evidence synthesis is a growing field which promises to reduce the human effort required to produce systematic reviews and other forms of evidence synthesis. A major strand of the literature works on screening, where several papers have developed or evaluated active learning. Active learning is an iterative process where documents screened by humans are used to train a machine learning model to predict the relevance of unseen papers. The most relevant studies are rated by a human, which generates more labels to feed back to the machine. By prioritising those studies most likely to be relevant, a human reviewer most often identifies all relevant studies - or above a given threshold proportion of relevant studies - before having seen all the documents in the corpus. The documents not yet seen by the human when they reach the given recall threshold have been represented as the work saved.

In live review settings, however, the recall - or proportion of relevant studies screened - remains an unknown until all documents have been screened. In order for work to really be saved, reviewers have to stop screening while uncertain about the recall. The lack of appropriate stopping criteria has been identified as a research gap [1], and existing approaches have fallen into the following categories

- **Sampling criteria:** Reviewers estimate the number of relevant documents by taking a random sample at the start of the process. They stop when this number, or a given proportion of it, has been reached [2]
- **Heuristics:** Reviewers stop when a given number of irrelevant articles are seen in a row [3].
- **Pragmatic criteria:** Reviewers stop when they run out of time [4].

We show in this paper how first two criteria are inadequate. This lack of adequate criteria challenges users of AL systems in live reviews, as authors are uncertain

when to stop, and have no reliable way of reporting when they stopped and what that may say about the recall they reached.

This study proposes a system for estimating the recall based on random sampling of remaining documents. We draw on the literature on binomial distributions [5], to illustrate how AL users can predefine a threshold in terms of uncertainty and recall, and use this to transparently save work with machine learning, while making a statement like "There is a <5% chance that we achieve a recall under 95%".

We evaluate this stopping criteria on real-world systematic review datasets on which active learning systems have previously been tested.

Methods

Existing Stopping Criteria for Active Learning

We start by explaining the sampling and heuristic based stopping criteria before showing with toy data how the criteria falls short. Then we introduce our own suggested stopping criteria, and show its benefits with toy data, before testing all criteria on real world datasets.

Sampling Based Stopping Criteria

The stopping criterion suggested by [2] involves establishing the Baseline Inclusion Rate (BIR), by taking a random sample at the beginning of screening. This is used to estimate the number of relevant documents in the whole dataset. Reviewers continue to screen until this number, or a proportion of it corresponding to the desired level of recall, is reached.

However, the estimation of the BIR fails to take into account the sampling uncertainty. Figure 1.a shows for 50 random samples of 2,000 documents from 20,000 documents (where 5% of documents are relevant), the predicted number of documents after each document drawn. Depending on the luck of the draw, one might estimate the true number of documents as being

- much higher than the true value, meaning that the stopping criterion would never be reached - and no work could be saved
- much lower than the true value, meaning that the stopping criterion would be reached before the desired level of recall was actually achieved

We can estimate the uncertainty of a binomial outcome using an Agresti-Coull confidence interval (which has been shown to be conservative at levels near 0 [5]). Figure 1.b shows the 95% confidence interval (brown) and its upper bound (green) in the same set of random draws. Where the green line is above the above the lower (upper) grey line, identification of 100% (95%) of the estimated number of studies would be impossible, and no work would be saved.

Figure 2 explores different scenarios of true relevance and probability thresholds, showing that the higher the true proportion of relevant documents, the fewer documents one would have to sample to make the identification of a given fraction (the true number of relevant results divided by the upper estimate) possible. Even with a relatively high proportion of relevant documents, and a sample of 500, all relevant documents would have to be identified to have a 95% likelihood of achieving 95% recall. With low proportions of relevant documents, only an 80% recall can be estimated with confidence, even after a sample of 2000, and subsequently identifying all

is is less relevant here, back to literature and e if another is better

hy does it behave like is, is there a theoretical maximum for each? mething to do with absolute values? a Bug? Is certainty a function of ing near to 0?

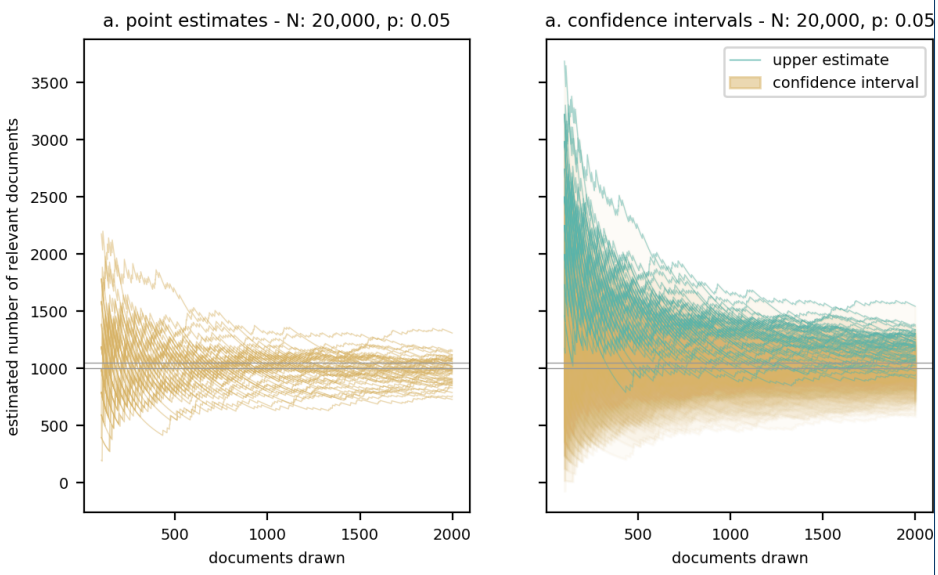

Figure 1 Uncertainty in estimating the baseline inclusion rate. a. shows the estimated BIR after drawing each document, for 50 random samples of from 20,000 documents, after each document is drawn. b. Shows the 95% Confidence interval (brown) and the upper bound (green) for each of the same samples. The two lines correspond to 100% of relevant documents, and $\frac{100}{0.95}$% of relevant documents.

relevant documents. N.B, these results ignore the random element of sample order, which explains the variation on figure 1. They show upper confidence bounds for accurate estimates of relevance.

Heuristic Based Stopping Criteria

Some studies give the example of heuristic based stopping criteria based on drawing a given number of irrelevant articles in a row [3] [+REFS]. We take this as a proxy for estimating that the proportion of documents remaining in the unseen documents is low. We find this a promising intuition, but argue that 1) it ignores uncertainty, as discussed in relation to the previous method; and 2) it misunderstands the significance of a low proportion of relevant documents in estimating the recall.

Figure 3 illustrates this second point. We show two scenarios with identical low proportions of relevant documents (note that estimating a proportion lower than 0.02 with 95% confidence requires 160 consecutive irrelevant documents). Recall is not simply a function of the relevance of unseen documents, but also of the number of unseen documents. This also means that where machine learning has performed well (as in the top figure), low proportions of irrelevant documents in those that remain are indicative of lower recall than where ML has performed less well.

demonstrate?

A Statistical Stopping Criterion for Active Learning

Figure 4 shows a workflow for the approach proposed in this paper. The random sampling occurs at the end of the process, and is used to estimate the number of relevant documents remaining and the total number of relevant documents. The

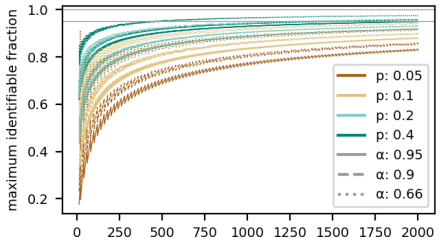

Figure 2 Maximum indentifiable fraction of relevant results under different levels of true relevance p and confidence thresholds α. Maximum identifiable fraction is the true number of relevant documents divided by the upper estimate.

Parameter	Description	Estimation
N	Total number of studies	*Observed*
S	Number of studies coded by humans	*Observed*
U	Number of studies not yet coded by humans	*Observed* $(N - S)$
p_S	relevance of documents rated by humans	*Observed*
α	Acceptable uncertainty level	*Given*
κ		
n	Number of randomly drawn documents	*Observed*
$\tilde{n}$		
κ		
$\hat{R}$	Estimated recall	
$\dot{R}$	Minimum estimated recall at a given confidence threshold	

Table 1 Parameters

rationale is to limit the uncertainty to a subsection of the dataset: that which has not yet been screened. As reviewers continue to draw random documents, the uncertainty range decreases, and the proportion of the data about which one is uncertain also decreases.

Table 1 shows the parameters, known, estimated, and given, available during the random sampling process and required to estimate recall. Expanding on what was stated before, recall - at a given confidence threshold is a function of 1) the upper estimate of the relevance of remaining documents, 2) The estimated relevance of all documents in the dataset, 3) The proportion of documents not yet seen.

Figure 5 shows the minimum estimated recall for a set of confidence intervals, along with the actual recall (in grey) for a case where 800 out of 2,000 documents have been reviewed, 2% of remaining documents are relevant, and 20% of all documents are relevant. We see that after 95% recall has actually been achieved, but before 100% of documents have been been seen, we can be confident at each of the given confidence levels that 95% of relevant documents habe identified. All estimates of recall only reach 100% after all documents have been seen, as it is not possible to exclude the possibility, at any given confidence level, that the proportion of relevant documents is greater than 0 . Figure 6 shows the various trajectories of a 95% minimum recall (blue) and actual recall (grey) given the same paramaters for 200 random samples.

ctually maybe this is
ssible if we work with
hole numbers? How
es that change things?

That the blue lines consistently meet a 95% threshold to the right of the grey lines indicates that reviewers have to see more than 95% of relevant documents in order to be confident that they have achieved their threshold. We call the number of documents it is necessary to review to establish with confidence that reviewers

have achieved 95% recall after they have already achieved 95% recall the *additional burden*.

In figure [], we investigate how the total number of documents and the proportion of relevant documents affect the additional burden .

Use the number of documents seen instead of p

Evaluation

Evaluation Data

Results

Discussion

Conclusion

Competing interests
The authors declare that they have no competing interests.

Author's contributions
Text for this section ...

Acknowledgements
Max Callaghan is supported by a PhD scholarship from the Heinrich Böll Foundation

Author details
[1]Mercator Research Institute on Global Commons and Climate Change, Torgauer Straße, 10829 Berlin, Germany.
[2]Priestley International Centre for Climate, University of Leeds, Leeds , LS2 9JT Leeds, United Kingdom.

References

1. Alexandra Bannach-Brown, Piotr Przybyła, James Thomas, Andrew S. C. Rice, Sophia Ananiadou, Jing Liao, and Malcolm Robert Macleod. Machine learning algorithms for systematic review: reducing workload in a preclinical review of animal studies and reducing human screening error. *Systematic Reviews*, 8(1):1–12, 2019.
2. Ian Shemilt, Antonia Simon, Gareth J. Hollands, Theresa M. Marteau, David Ogilvie, Alison O'Mara-Eves, Michael P. Kelly, and James Thomas. Pinpointing needles in giant haystacks: Use of text mining to reduce impractical screening workload in extremely large scoping reviews. *Research Synthesis Methods*, 5(1):31–49, 2014.
3. Piotr Przybyła, Austin J. Brockmeier, Georgios Kontonatsios, Marie Annick Le Pogam, John McNaught, Erik von Elm, Kay Nolan, and Sophia Ananiadou. Prioritising references for systematic reviews with RobotAnalyst: A user study. *Research Synthesis Methods*, 9(3):470–488, 2018.
4. Makoto Miwa, James Thomas, Alison O'Mara-Eves, and Sophia Ananiadou. Reducing systematic review workload through certainty-based screening. *Journal of Biomedical Informatics*, 51:242–253, 2014.
5. Lawrence D. Brown, Tony T. Cai, and Anirban DasGupta. Interval estimation for a Binomial Proportion. *Statistical Science*, 16(2):101–133, 2001.

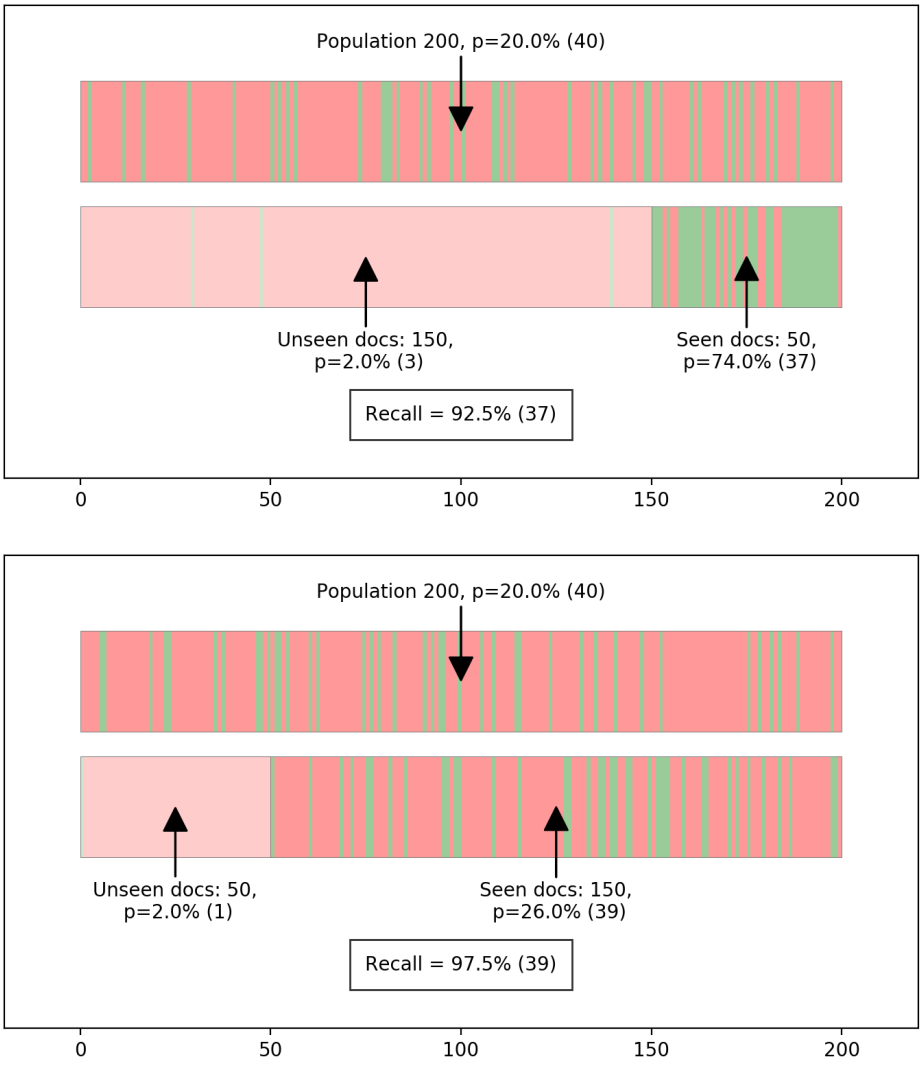

Figure 3 Similar low proportions of relevant documents in unseen documents with different consequences for recall. The top bar shows a random distribution of relevant documents (green) and irrelevant documents (red) at a given proportion of relevance. The bottom bar shows distributions of relevant and irrelevant documents in hypothetical sets of seen (right) and unseen (left - transparent) documents.

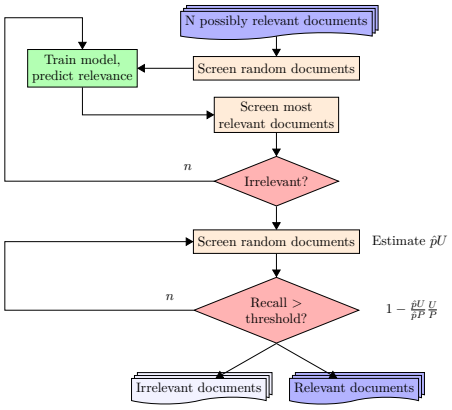

Figure 4 A workflow for active learning in screening with a statistical stopping criterion

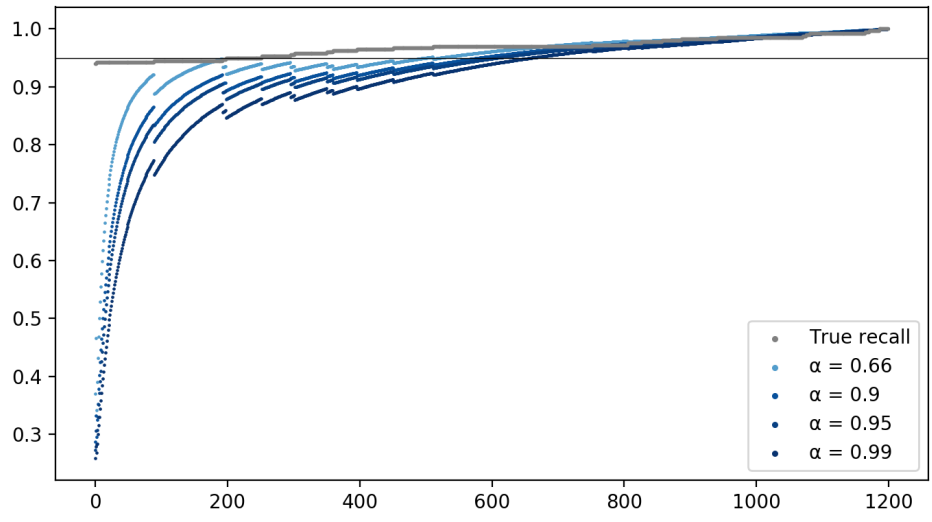

Figure 5 Sample recall

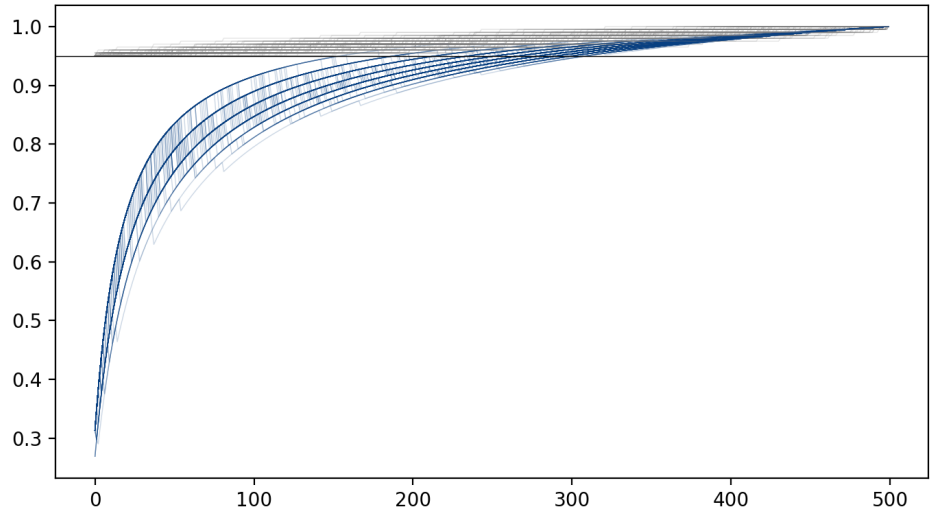

Figure 6 Sample recall variation

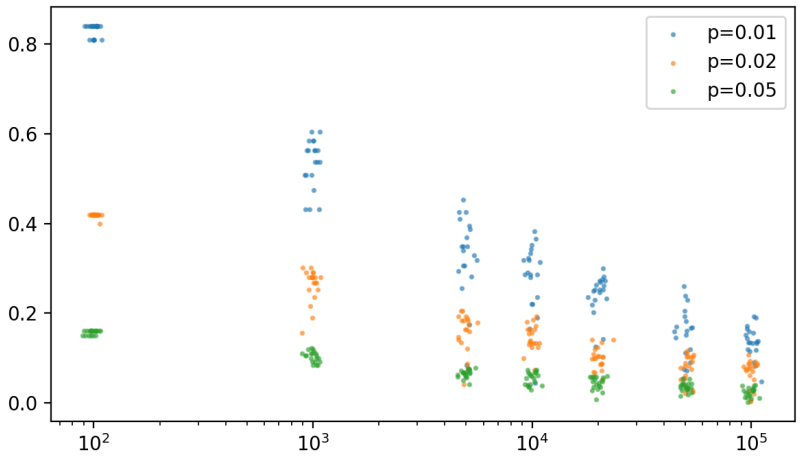

Figure 7 Additional burden for 20 random samples where the 95% threshold has been reached, and the remaining documents are 1%, 2% or 5% relevant, for different total numbers of documents.