Data Balancing for Technologically Assisted Reviews: Undersampling or Reweighting

Zhe Yu and Tim Menzies[1]

NC State University, Raleigh NC 27695, USA,
zyu9@ncsu.edu,tim.menzies@gmail.com

Abstract. This paper provides approaches for automated support of citation screening in systematic reviews. Continuous active learning is chosen as our baseline approach, above which, two data balancing techniques are applied to handle the imbalance problem. These two techniques, aggressive undersampling and reweighting are tested and compared on 20 data sets for Diagnostic Test Accuracy (DTA) reviews. Results are evaluated by **last_rel** and suggest that reweighting outperforms undersampling as it not only balances the training data, but also *emphasizes* the "content relevant" examples over "abstract relevant" ones and thus helps to retrieve "content relevant" papers earlier.

Keywords: technologically assisted reviews, active learning, data balancing

1 Introduction

This paper is a participant working note for the task of technologically assisted reviews in empirical medicine [7] in CLEF eHealth 2017 [6]. This task is about applying machine learning techniques to facilitate medical researchers conducting systematic reviews. More specifically, the task focuses on Diagnostic Test Accuracy (DTA) reviews since search in this area is generally considered the hardest, and a breakthrough in this field would likely be applicable to other areas as well [7]. Twenty DTA reviews data sets are provided for training and thirty for testing. The problem statement of this task is:

> *Given the results of a Boolean Search how to make Abstract and Title Screening more effective.*

Here, in this paper, we further specify our problem to be:

> *Screen least amount of papers to retrieve most (or all) relevant ones.*

This leads directly to the evaluation method– **last_rel** [7], which measures the number of documents need to be screened before retrieving all relevant documents.

Previously, we analyzed the equivalent problem in software engineering (SE) and built a high performing method FASTREAD that combines a wide range of techniques taken from from electronic discovery and evidence based medicine [12]. Those results suggested that FASTREAD, which took aggressive undersampling from patient active learning [11, 10] and the rest from continuous active learning [2–4], outperforms both of the original algorithms on SE reviews data [12]. It indicated that, at least on SE reviews data, continuous active learning is an efficient approach, and data balancing can further improve its performance.

While the above results are promising, we advise against applying the conclusions directly to the empirical medicine task since the target corpus are very different (one from SE reviews and one from DTA reviews). In addition, the DTA reviews data have two levels of query results, one from title and abstract screening, the other from document screening, while the SE reviews data [12] only have the query results from document screening. That is, we feel that when properly considered, reweighting can be another way to balance the training data while more weights are put on papers identified as "content relevant" over those identified as only "abstract relevant" or "not relevant". In this way, reweighting not only balances the two classes, but also favors "content relevant" examples when training the model.

Besides the two-level query results, DTA reviews data also offer a brief description of the topic being screened, which could be a great source for "Auto-Syn" described in [13] and [3]. Utilizing the description as an initial seed training example would provide better chance to retrieve "relevant" papers earlier and reduce variances in the experiments (comparing to a random start-up). Note that in order to train a classifier on just one "relevant" example (the description of the topic), *Presumptive non-relevant examples* are generated [3]. This technique randomly samples from the unlabeled examples and treats the sampled examples as "not relevant" in training. The low prevalence of "relevant" examples makes this technique reasonable.

The rest of the paper provides details about different approaches tested on training data and analyzes the results. Numerous engineering decisions have been made without fully tested due to limited time. Followed by conclusions and future works at last.

2 Method

In this section, we provide details on three approaches:

- **CAL:** a baseline approach from Cormack et al. [2–4].
- **AU:** add data balancing method called aggressive undersampling [11, 10] to the baseline approach CAL.
- **RW:** add reweighting method ("content relevant" papers weight more than other papers in training) to the baseline approach CAL.

2.1 Baseline: CAL

Besides the overall framework as continuous active learning [2–4], the baseline approach applies several predefined engineering decisions same as our previous work [12]. The entire work flow can be described as follows:

1. **Corpus collection:** collect titles and abstracts of papers in search results.
2. **Auto-Syn:** add the topic description into the corpus and label it as "abstract relevant".
3. **Preprocessing:** stemming, stop words removal, bag of words.
4. **Featurization:** term frequency, feature selection by tf-idf score (to 4000 terms), l2 normalization.
5. **Training:** train a binary classifier (linear SVM) on all the labeled papers, "content relevant" and "abstract relevant" papers are treated as one class– "relevant" while "not relevant" papers are the other class in the training. *Presumptive non-relevant examples* are generated to enrich the "not relevant" class examples.
6. **Certainty sampling:** use the trained classifier to predict on the rest unlabeled papers. Sample $N = 10$ papers with highest probability to be "relevant" according to the classifier.
7. **Review[1]:** ask reviewers to review the sampled papers by titles and abstracts, label each as "abstract relevant" or "not relevant". For those papers labeled as "abstract relevant", reviewers are asked to further review on content and decide whether to label each as "content relevant". Go back to 5 until stop rule is satisfied (every "content relevant" paper has been retrieved).

2.2 Aggressive Undersampling: AU

First proposed by Wallace et al. in 2010 [11], aggressive undersampling is a technique applied in patient active learning to balance the training classes. The only difference between **AU** and the baseline approach **CAL** is:

> 5. **Training:** train a binary classifier (linear SVM) on all the labeled papers, "content relevant" and "abstract relevant" papers are treated as one class– "relevant" while "not relevant" papers are the other class in the training. *Presumptive non-relevant examples* are generated to enrich the "not relevant" class examples. **If there are more than $M = 30$ "relevant" papers in the training set, aggressive undersampling is performed. It undersamples the "not relevant" papers to the same number as "relevant" ones by throwing away the "not relevant" papers closes to the SVM decision hyperplane. After aggressive undersampling, SVM is retrained on the balanced training data.**

The threshold of $M = 30$ is applied to avoid training an SVM model on too few papers [12].

2.3 Reweighting: RW

Reweighting (RW) is a new approach which takes advantage of the two-level labels offered by DTA reviews data. The difference between **RW** and the baseline approach **CAL** is:

[1] The actual experiments are carried out without real human reviewers. When asked for labels, the true labels in the data sets are queried instead of a human reviewer.

5. **Training:** train a binary classifier (linear SVM) on all the labeled papers, "content relevant" and "abstract relevant" papers are treated as one class– "relevant" **but "content relevant" papers have** $W =$ 10 **times the weight of "abstract relevant" or "not relevant" ones.** *Presumptive non-relevant examples* are generated to enrich the "not relevant" class examples.

The reweighting parameter of $W = 10$ is chosen quite arbitrarily without fully tested due to the limited time.

3 Experiment

Experiments are conducted in a "pseudo" way following the procedures in Section 2. When a paper is asked to be reviewed, its true label is queried without any real human review process. As a result, the experiments become repeatable and reproducible.

3.1 Data

Twenty data sets on DTA reviews are provided as training sets for the task of technologically assisted reviews in empirical medicine [7]. These data sets provide two-level query results, one for title and abstract screening and one for content screening. As a result, we label each paper in the data sets as one of the three classes:

- **Not relevant:** papers excluded by title and abstract screening.

Table 1. Descriptive statistics for experimental data sets.

	Content	Abstract	Total		Content	Abstract	Total
Topic1	2	30	3241	Topic35	9	98	3857
Topic4	28	442	8180	Topic37	12	154	1576
Topic6	2	6	15078	Topic38	5	109	12704
Topic9	60	98	1162	Topic43	27	48	43335
Topic11	8	59	1457	Topic44	30	206	3149
Topic14	20	63	14907	Topic45	1	42	316
Topic19	1	1	12704	Topic50	41	143	7990
Topic23	48	200	1938	Topic53	19	67	1310
Topic28	3	5	3964	Topic54	14	27	1499
Topic33	60	604	8186	Topic55	45	92	2542

"Content" column displays the number of "content relevant" papers; "Abstract" column displays the number of "content relevant" papers plus the number of "abstract relevant" papers; "Total" column displays the total number of papers. Topic 1, 6, 19, 28, and 45 (colored in red) are considered "not good" for **last_rel** evaluation due to their lack of "content relevant" papers (fewer than 5).

- **Abstract relevant:** papers included by title and abstract screening but excluded by content screening.
- **Content relevant:** papers included by title and abstract screening and content screening.

Statistics for the twenty data sets are presented in Table 1 where five sets are considered to be "not good" for **last_rel** evaluations. The reason behind is that pure "luck" might affect the result when the target is to retrieve the only 1 (or 2, or 3) "content relevant" paper.

3.2 Performance Metrics

Since the objective is to screen least amount of papers to retrieve most (or all) relevant ones, we choose **last_rel** for evaluation. More specifically, we use the number of papers screened when every "content relevant" one is retrieved as the performance score to take advantage of the two-level labels offered by DTA reviews data. This makes our **last_rel** metrics different from that used in [7].

The lower the **last_rel** score is, the fewer papers need to be manually screened, thus the better performance. To capture the possible variances, experiments of each method on every data set (topic) is repeated 10 times with different random seeds (which affect the *presumptive non-relevant examples* generated and thus introduce variances). The **last_rel** score for each repeat is collected while medians and iqrs (75th-25th percentile) are calculated for comparison. Scott-Knott [9] analyses are applied on each topic to rank the performance of each treatment. Since the **last_rel** scores are in asymmetric and non-normal distributions, Cliff's Delta [1] and bootstrapping [5] are applied for non-parametric hypothesis test; i.e. two treatments are ranked differently in Scott-Knott analysis if both bootstrapping and the effect size test agreed that the division is statistically significant (99% confidence) and not a small effect (Cliff's Delta $\geq$ 0.147).

3.3 Results

Table 2 shows the results on 20 topics from the training set. The first thing we notice is that there is no treatment ranks highest (colored in green) across every topic. One treatment may outperform others in one topic but performs poorly in another topic. In addition, no domination can be found among the three treatments (we say treatment A dominates treatment B if A performs consistently better than B across all topics).

Therefore, when it comes to the question of which treatment is the best, it really depends on the data. However, we did summarize the results in Table 2 and count the number of "wins" and "losts" of each treatment. As shown in Table 3, statistically, **reweighting (RW)** wins more and loses less than any other treatment. As a result, among these three treatments, we recommend reweighting (RW), which over-weights the "content relevant" examples to balance training data as well as emphasize "content relevant" examples.

Table 2. Experimental Results.

	MEDIAN			IQR		
	RW	AU	CAL	RW	AU	CAL
Topic1	510	890	885	205	7	10
Topic4	260	410	385	70	62	122
Topic6	5475	12270	6055	327	7	440
Topic9	690	750	690	0	0	0
Topic11	75	90	80	17	10	7
Topic14	110	115	110	10	25	17
Topic19	8320	6160	8320	0	0	7
Topic23	920	840	1040	0	27	0
Topic28	1715	1525	1600	17	27	17
Topic33	4360	3780	4970	0	62	0
Topic35	210	260	405	27	10	115
Topic37	310	380	475	20	27	35
Topic38	490	960	980	15	447	97
Topic43	180	1140	230	37	210	17
Topic44	670	510	945	92	37	50
Topic45	20	10	10	15	7	7
Topic50	425	445	535	65	35	105
Topic53	340	620	280	0	60	0
Topic54	510	440	440	10	0	0
Topic55	740	850	610	0	17	0

Results collected from 10 repeated runs on 20 topics. Both medians and iqrs are lower the better. For each topic, aggressive undersampling (AU) and reweighting (RW) are compared along with the baseline method continuous active learning without data balancing (CAL). Scott-Knott analyses (with Cliff's Delta and bootstrapping for non-parametric hypothesis test) are applied to rank each treatment. The treatments with highest rank are colored in green while the treatments with lower ranks than the baseline (CAL) are colored in gray.

Another gain from these experiments is that **data balancing** techniques do improve the performances. As indicated in Table 2, on 19 out of 20 (or 14 out of 15) topics, reweighting (RW) or aggressive undersampling (AU) ranks highest; on 13 out of 20 (or 9 out of 15) topics RW or AU ranks higher than continuous active learning (CAL) without data balancing. This also suggests that the ensemble of RW and AU to leverage the advantages from both data balancing techniques might offer even better results. We plan to explore this in our future works.

Variances are within an acceptably low range (except for some of the "not good" topics) thanks to "Auto-Syn" technique. Therefore the results are considered to be stable and repeatable.

Table 3. Summary of the Experimental Results.

	In all 20 topics		In 15 “good” topics	
	Top Rank	Lower Rank than Baseline	Top Rank	Lower Rank than Baseline
RW	14	3	11	2
AU	9	6	6	5
CAL	7	NA	6	NA

“Top Rank” column displays the number of times one treatment ranks highest while “Lower Rank thanBaseline” column displays the number of times one treatment ranks lower than baseline treatment (CAL). The first two columns count all 20 topics while the last two columns only count “good” topics (excluding topics colored in red in Table 1 and 2). One treatment is considered better than another if the number in “Top Rank” is larger while the number in “Lower Rank thanBaseline” is smaller.

4 Conclusion

How to retrieve most (or all) relevant documents by screening least amount of the candidate ones is a difficult problem which is also known in the Information Retrieval (IR) domain as the total recall problem. Proposed by Cormack et al. in 2014, continuous active learning has been an excellent algorithm to solve the problem [2–4]. It was also adopted as a baseline method in the total recall task of TREC 2015 [8]. This work extended continuous active learning method by testing two different data balancing techniques. Experimental results suggested that there were no single treatment that outperforms any other treatment across all topics. However, statistically, reweighting (RW) was considered to be most powerful for the total recall task. This treatment applied “Auto-Syn” with topic description as seed training data, generated “presumptive non-relevant examples” before training to enrich the “not relevant” class, over-weighted the “content relevant” examples for data balancing. With the reweighting treatment, training examples were balanced (thus the model will not over-fit on “not relevant” class), and the model was trained to “favor” the “content relevant” examples which had a positive effect on retrieving every “content relevant” paper earlier.

Due to the limited time, only one aspect (data balancing) has been explored in this study. This does not imply that other aspects of the total recall task are not worth exploring. The plans of future work include:

- Explore the ensemble of reweighting and aggressive undersampling and other possible data balancing techniques.
- Many parameters in the tested treatments are chosen quite arbitrarily. Parameter tuning can be applied to see if these parameters affect the conclusion and whether we can find a better set of parameters.

- Different featurization techniques can be applied to extract “richer” features than bag-of-words or term frequencies; e.g. word vectors and citation link features might be useful for measurement of relevance.
- Human errors can be injected to test how robust the active learning methods are and to what level of error rate can the system perform normally.

References

1. Cliff, N.: Dominance statistics: Ordinal analyses to answer ordinal questions. Psychological Bulletin 114(3), 494 (1993)
2. Cormack, G.V., Grossman, M.R.: Evaluation of machine-learning protocols for technology-assisted review in electronic discovery. In: Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval. pp. 153–162. ACM (2014)
3. Cormack, G.V., Grossman, M.R.: Autonomy and reliability of continuous active learning for technology-assisted review. arXiv preprint arXiv:1504.06868 (2015)
4. Cormack, G.V., Grossman, M.R.: Scalability of continuous active learning for reliable high-recall text classification. In: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. pp. 1039–1048. ACM (2016)
5. Efron, B., Tibshirani, R.J.: An introduction to the bootstrap. CRC press (1994)
6. Goeuriot, L., Kelly, L., Suominen, H., Névéol, A., Robert, A., Kanoulas, E., Spijker, R., Palotti, J.R.M., Zuccon, G.: Clef 2017 ehealth evaluation lab overview. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction - 8th International Conference of the CLEF Association, CLEF 2017, Dublin, Ireland, September 11-14, 2017, Proceedings. Lecture Notes in Computer Science, Springer (2017)
7. Kanoulas, E., Li, D., Azzopardi, L., Spijker, R.: Overview of the CLEF technologically assisted reviews in empirical medicine. In: Working Notes of CLEF 2017 - Conference and Labs of the Evaluation forum, Dublin, Ireland, September 11-14, 2017. CEUR Workshop Proceedings, CEUR-WS.org (2017)
8. Roegiest, A., Cormack, G.V., Grossman, M., Clarke, C.: Trec 2015 total recall track overview. Proc. TREC-2015 (2015)
9. Scott, A.J., Knott, M.: A cluster analysis method for grouping means in the analysis of variance. Biometrics pp. 507–512 (1974)
10. Wallace, B.C., Small, K., Brodley, C.E., Trikalinos, T.A.: Active learning for biomedical citation screening. In: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 173–182. ACM (2010)
11. Wallace, B.C., Trikalinos, T.A., Lau, J., Brodley, C., Schmid, C.H.: Semi-automated screening of biomedical citations for systematic reviews. BMC bioinformatics 11(1), 1 (2010)
12. Yu, Z., Kraft, N.A., Menzies, T.: How to read less: Better machine assisted reading methods for systematic literature reviews. CoRR abs/1612.03224 (2016), http://arxiv.org/abs/1612.03224
13. Zhang, H., Lin, W., Wang, Y., Clarke, C.L., Smucker, M.D.: Waterlooclarke: Trec 2015 total recall track. In: TREC (2015)