Assessor Disagreement and Text Classifier Accuracy

William Webber
College of Information Studies
University of Maryland
United States of America
wew@umd.edu

Jeremy Pickens
Catalyst Repository Systems
Denver, CO
United States of America
jpickens@catalystsecure.com

ABSTRACT

Text classifiers are frequently used for high-yield retrieval from large corpora, such as in e-discovery. The classifier is trained by annotating example documents for relevance. These examples may, however, be assessed by people other than those whose conception of relevance is authoritative. In this paper, we examine the impact that disagreement between actual and authoritative assessor has upon classifier effectiveness, when evaluated against the authoritative conception. We find that using alternative assessors leads to a significant decrease in binary classification quality, though less so ranking quality. A ranking consumer would have to go on average 25% deeper in the ranking produced by alternative-assessor training to achieve the same yield as for authoritative-assessor training.

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Systems and software—*performance evaluation*.

General Terms

Evaluation

Keywords

Text classification, evaluation, assessor disagreement

1. INTRODUCTION

Text classification based upon machine learning is a useful tool for text retrieval tasks on corpora with many relevant documents, where high recall is required, and where the searcher is willing to devote significant effort to the task. One such environment is that of e-discovery—the retrieval of responsive documents in civil law—and classification technologies have been widely deployed there.

To learn a relevance model, a machine learner is provided with example documents, annotated by a human assessor. The assessor making the relevance judgments may not, however, be the person whose conception of relevance is authoritative. In e-discovery, for instance, senior lawyers commonly delegate assessment to junior lawyers or contract paralegals, due to time and cost constraints.

SIGIR'13, July 28–August 1, 2013, Dublin, Ireland.

Human assessors frequently disagree on document relevance [8], which questions the use of non-authoritative assessors to train text classifiers. How reliable are classifiers trained by non-authoritative assessors when evaluated by the authoritative conception of relevance? Does the classifier compensate for the disagreement between assessors, or does it amplify it?

2. PREVIOUS WORK

In an experiment reported by Voorhees [8], TREC AdHoc documents were assessed by two alternative assessors, and high levels of assessor disagreement were observed. Based on simulation experiments, Carterette and Soboroff [2] find that overly-conservative assessors (those who find fewer documents relevant) distort retrieval effectiveness evaluation less than liberal ones do.

In the e-discovery domain, Grossman and Cormack [5] compare non-authoritative assessors with automated techniques guided by authoritative feedback, finding the latter to be at least as reliable as the former when evaluated against the authoritative conception of relevance. Webber [9] analyses assessor agreement levels on the same dataset, finding considerable variability in assessor reliability.

Brodley and Friedl [1] present methods for automatically identifying mislabeled training data by using ensemble classifiers to detect outliers. Ramakrishnan et al. [7] similarly use a Bayesian network to detect outliers in textual data. Such methods do not work, however, if the annotator is consistently incorrect.

3. MATERIALS AND METHODS

We distinguish two assessors: the training assessor, who makes the annotations of the training examples; and the testing assessor, whose conception of relevance the output classifier is intended to represent. Where training and test assessor are the same, we refer to the task as self-classification. Where the assessors are different, we refer to the task as cross-classification.

3.1 Metrics

We use F_1 score—the harmonic mean of precision and recall—as our measure of effectiveness for binary classification. To measure ranking quality, we calculate the maximum F_1 score achievable across all possible cutoff points in the ranking (termed hypothetical F_1 by Cormack et al. [4]). Area under the ROC curve gives similar trends to those reported here. Significance testing is by paired two-tailed t tests.

3.2 Dataset

Our dataset is taken from the TREC 4 AdHoc track. In that year, the organizers arranged for selected documents to be triply-assessed, first by the author of the TREC topic, and then by two additional assessors, who were authors of other TREC topics [8].

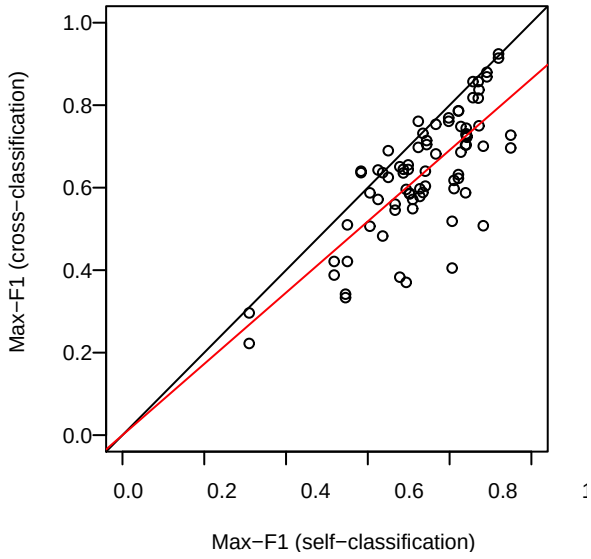

Figure 1: (Ranking) Maximum-F1 score for cross classification versus self-classification, with the original assessor as target assessor. An origin-anchored regression line is drawn.

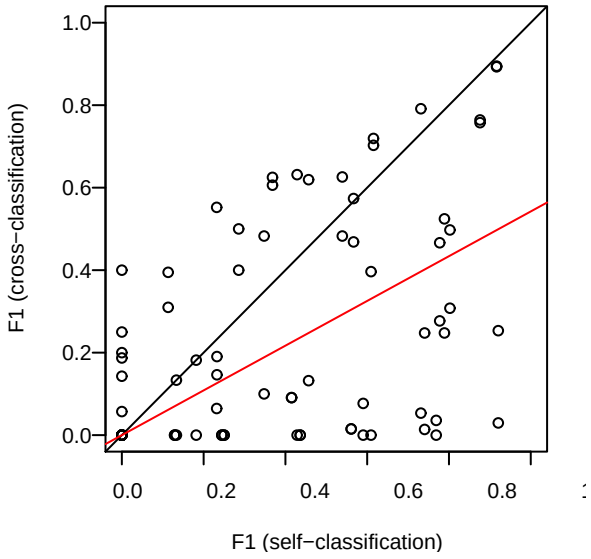

Figure 2: (Binary) F1 score for cross classification versus self-classification.

We treat the original assessor as the authoritative, testing assessor, and separately treat each additional assessor as a training assessor for cross-classification.

We restrict our document set to the Associated Press (AP) subcollection, in order to avoid certain biases in the original (non-random) selection of documents for multiple assessment. We include only those 39 (of the original 49) topics for which all three assessors found at least 8 AP documents relevant. The mean number of relevant documents per topic is 73 (standard deviation 60), and of irrelevant documents 191 (sd 31). The mean F_1 between the original and alternative assessors is 0.63 (sd 0.21).

3.3 Classifier

We use LibSVM as our classifier [3], with a linear kernel and default settings. Features are term TF*IDF scores, using length normalization, the Lovins stemmer, case folding, and stop word removal. Inverse document frequency was calculated only on AP documents multiply-assessed for at least one topic.

As the dataset is small, classification is approximated by classifying each tenth of the collection using a model trained on the other nine-tenths. The tested tenths are then amalgamated to form a single, margin-based ranking. Holdout experiments showed a mean Kendall's τ of 0.88 between document rankings produced by different fold models, indicating high stability between models.

LibSVM optimizes its binary classification for accuracy, but this proved to give poor results for the F_1 measure. Instead, we create binary rankings by fitting probabilities using the method of Platt [6], then choosing the cutoff point that optimizes F_1 in expectation.

4. RESULTS

4.1 Self- versus cross-classification

We begin by comparing the ranking effectiveness of the classifier trained by the authoritative assessor (self-classification) with that trained by an alternative assessor (cross-classification). Figure 1 compares the max-F_1 scores achieved by the two approaches.

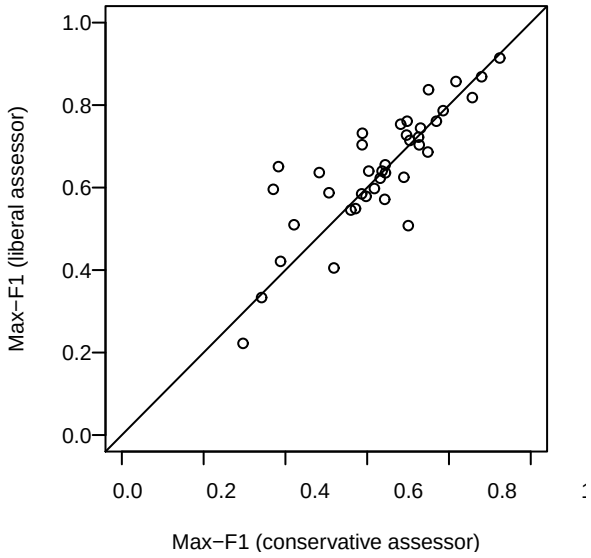

Figure 3: Cross-classification effectiveness of conservative versus liberal alternative assessor, with original assessor as target, as measured by maximum-F1 score.

Mean max-F_1 is 0.738 for self- and 0.637 for cross-classification; the difference is highly significant ($p < 0.0001$). Cross-classification leads to an average max-F_1 score 14% than self-classification.

Next we consider binary classification, as shown in Figure 2. Mean binary F_1 is 0.629 for self- and 0.456 for cross-classification, again a highly significant difference. Cross-classification leads to a 28% lower F_1 score than self-classification, a greater fall than for max-F_1. Cross-classification seems to harm selection of a binary cutoff even more than it does ranking of the documents.

4.2 Comparing different assessor types

An interesting question is whether, given an assessor disagrees

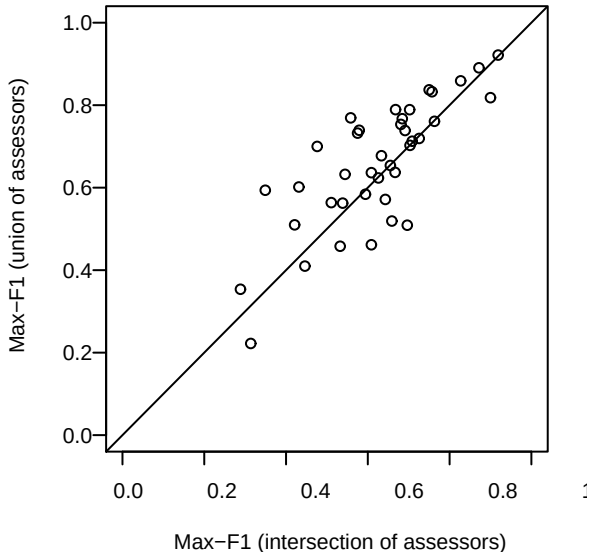

Figure 4: Cross-classification effectiveness of the union of alternative assessors' relevant documents versus the intersection, measured using maximum-F1 score.

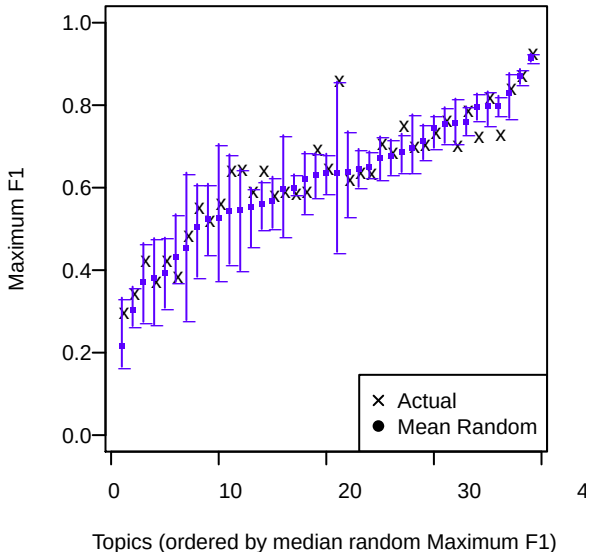

Figure 5: Actual cross-classification effectiveness versus range of effectivenesses of randomly-degraded cross-classification, measured using maximum-F1 score. The mean and 95% intervals on the random cross-classifications are shown. Topics are sorted by mean random cross-classifier effectiveness. There are 81 random simulations of alternative assessors for each topic.

with the authoritative conception, it is better that the assessors tends to assign more documents as relevant (the assessor is liberal), or fewer (the assessor is conservative). We explore this question by denoting the alternative assessor with the lower prevalence for each topic the conservative assessor, and the assessor with the higher prevalence the liberal assessor. Figure 3 compares the max-F_1 scores on the rankings produced via cross-classification using conservative versus liberal assessors. Mean max-F_1 is 0.629 for the conservative assessors, 0.646 for the liberal ones. The difference, however, is not significant ($p > 0.1$).

A related question is how to combine multiple assessments, where available, when creating training data. Should the union of the documents found relevant by either assessor be marked relevant in the training data, or the intersection (that is, only documents both assessors find relevant)? Figure 4 compares two alternatives: marking as relevant documents found relevant by either assessor (union), versus only those found relevant by both (intersection). The intersection of the assessors gives a mean max-F_1 of 0.623, the union one of 0.657, with the difference being statistically significant ($p < 0.05$). It seems on balance better to give more, if noisier, examples of relevant documents than fewer, if cleaner, examples. (Only retaining examples on which both assessors agreed was also tried; the mean max-F_1 score is intermediate between that for the intersection and that for the union.)

4.3 Random disagreement

The previous sections have examined the absolute loss of effectiveness from using non-authoritative assessors to train the classifier. Is this loss greater or less than one would expected, given inter-assessor agreement? One way of answering this is to compare cross-classification effectiveness of the actual alternative assessor, with that of other randomly simulated alternative assessors having the same agreement level. We do this by starting with the original assessments and the false positive and false negative counts, FP and FN, of the alternative assessor (we arbitrarily choose the first alternative assessor for this experiment). We then random select FP of the originally irrelevant documents and mark them relevant, and FN of the originally relevant documents, and mark them irrelevant, creating a simulated alternative assessor training set. We then train a cross-classifier on this simulated set, and compare its effectiveness with the actual alternative assessor.

Figure 5 compares simulated cross-classification effectiveness (across 81 simulations per topic) with that of the actual alternative assessor, measured using max-F_1. The mean of the actual max-F_1 scores is 0.633, that of the median random 0.615; the difference is statistically significant ($p < 0.05$). On average, the actual alternative assessor gives slightly better ranking quality than inter-assessor agreement would predict, though the difference is small. There is considerable variability between topics (or assessors): actual is outside the empirical 95% interval for 7 of the 39 topics (above for 3, below for 4).

4.4 User effort

Differences in effectiveness have been expressed in previous sections in terms of the system evaluation metrics of F_1 and max-F_1. These results can be difficult to interpret in terms of the actual cost to the user of poorer performance. One way of measuring this cost is how much further down the ranking one must go in order to achieve a certain level—say 75%—of recall. In e-discovery, productions are often finalized by manually reviewing the ranking from the top down to the point where it is estimated that a certain threshold of recall (and 75% is one such threshold[1]) has been achieved, so depth to achieve 75% recall is a reasonable measure of one component of expense in e-discovery.

Figure 6 compares the proportion of the ranking that must be processed to achieve 75% recall for cross-classification with that

[1]See, for instance, *Global Aerospace Inc., et al., v. Landow Aviation, L.P., et al.*, No. CL 61040 (Va. Cir. Ct. Apr. 9, 2012) ("Memorandum in support of motion for protective order approving the use of predictive coding").

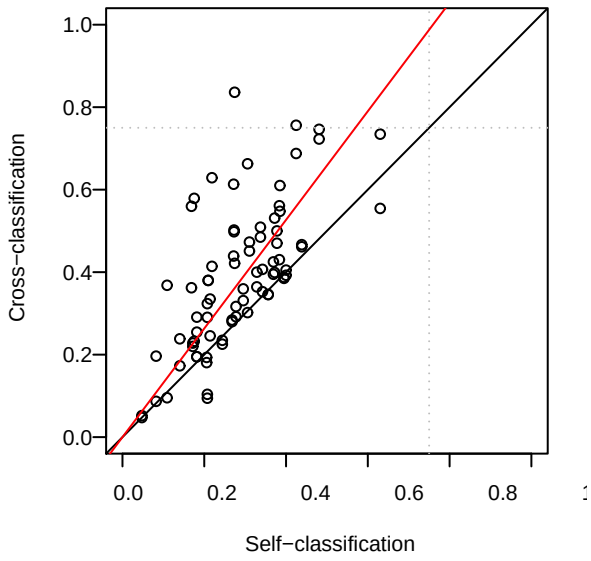

Figure 6: Proportion of ranking that must be processed in order to achieve 75% recall, under cross-classification and self-classification.

for self-classification. In the median case, using cross-classification requires that 24% more of the ranking must be processed than using self-classification, but around one in eight cases, processing must go to twice the depth or more.

5. CONCLUSION

In this paper, we have examined the loss of effectiveness that occurs when a text classifier is trained using annotations made by an assessor other than the authoritative assessor, whose conception of relevance is to be used to evaluate the classifier's effectiveness. We have found that using a non-authoritative assessor leads to a significant decrease in classifier reliability, of around 14% for ranking quality measured using maximum F_1, and twice that for binary classification measured using F_1 score. In terms of user effort, this means that around 24% more of the ranking must be processed to achieve recall of 75%. The liberality or conservativeness of the assessor does not make a significant difference to cross-classification reliability, though where multiple assessments are available, it seems slightly better to take the union of their relevance sets rather than their intersection as training data. Cross-classification leads to slightly better average performance than might be expected given the degree of inter-assessor disagreement (as measured via a random simulation experiment). However, for all of these findings, there is considerable variability between tasks and between assessors.

Considerable future work remains to be done. Though the training sets employed here have been sufficient to achieve creditable accuracy (mean F_1 of 0.629) on the low-yield ad-hoc tasks, larger training sets are used in many text-classification tasks, such as e-discovery, where a few thousand training examples are more common. Larger training sets may contain more redundancy, reducing the impact of assessor disagreement; though to the extent that disagreement is systematic rather than random, the reduction may be slight. Similarly, the relative desirability of liberal or conservative assessors, or of the union or intersection of multiple assessment sets, will likely be affected by the amount of training data. We have explored the question of user cost in terms of additional processing of the output ranking; another dimension of cost that a larger experimental training set would allow us to explore is the additional number of annotations required under cross-classification to achieve the same effectiveness as self-classification, and also what the (near-) maximum effectiveness achievable with both assessor types is. We intend to explore this question using TREC Legal Track data.

Finally, the evaluation metrics used in this paper include F1 and user effort. However, user effort does not always have uniform cost. One of the primary motivations for this work is that non-authoritative assessors (e.g. junior attorneys in an e-discovery matter) have a lower hourly cost than authoritative assessors (e.g. senior attorneys). One of the next phases of this research is integrating economic cost models with retrieval effectiveness metrics, to paint an overall picture of the cost of using non-authoritative, less accurate assessors.

Acknowledgments

This material is based in part upon work supported by the National Science Foundation under Grant No. 1065250. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

[1] Carla E. Brodley and Mark A. Friedl. Identifying mislabeled training data. *Journal of Artificial Intelligence Research*, 11:131–167, 1999.

[2] Ben Carterette and Ian Soboroff. The effect of assessor errors on IR system evaluation. In *Proc. 33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 539–546, Geneva, Switzerland, July 2010.

[3] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[4] Gordon V. Cormack, Maura R. Grossman, Bruce Hedin, and Douglas W. Oard. Overview of the TREC 2010 legal track. In Ellen Voorhees and Lori P. Buckland, editors, *Proc. 19th Text REtrieval Conference*, pages 1:2:1–45, Gaithersburg, Maryland, USA, November 2010.

[5] Maura R. Grossman and Gordon V. Cormack. Technology-assisted review in e-discovery can be more effective and more efficient than exhaustive manual review. *Richmond Journal of Law and Technology*, 17 (3):11:1–48, 2011.

[6] John C. Platt. Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. In Alexander J. Smola, Peter Bartlett, Bernhard Schölkopf, and Dale Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 61–74. MIT Press, 1999.

[7] Ganesh Ramakrishnan, Krishna Prasad Chitrapura, Raghu Krishnapuram, and Pushpak Bhattarcharyy. A model for handling approximate, noisy or incomplete labeling in text classification. In *Proc. 22nd International Conference on Machine Learning*, pages 681–688, Bonn, Germany, August 2005.

[8] Ellen Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing & Management*, 36(5):697–716, September 2000.

[9] William Webber. Re-examining the effectiveness of manual review. In *Proc. SIGIR Information Retrieval for E-Discovery Workshop*, pages 2:1–8, Beijing, China, July 2011.