Sequential Testing in Classifier Evaluation Yields Biased Estimates of Effectiveness

William Webber
Mossaab Bagdouri
iSchool, CS
University of Maryland
College Park, MD 20742, USA
{wew,mossaab}@umd.edu

David D. Lewis
David D. Lewis Consulting
Chicago, IL 60614, USA
sigir2013pap@
DavidDLewis.com

Douglas W. Oard
iSchool / UMIACS
University of Maryland
College Park, MD 20742, USA
oard@umd.edu

ABSTRACT

It is common to develop and validate classifiers through a process of repeated testing, with nested training and/or test sets of increasing size. We demonstrate in this paper that such repeated testing leads to biased estimates of classifier effectiveness. Experiments on a range of text classification tasks under three sequential testing frameworks show all three lead to optimistic estimates of effectiveness. We calculate empirical adjustments to unbias estimates on our data set, and identify directions for research that could lead to general techniques for avoiding bias while reducing labeling costs.

Categories and Subject Descriptors

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

General Terms

Performance

Keywords

Supervised learning, text categorization, evaluation, F-measure

1. INTRODUCTION

Classifiers are frequently produced by supervised learning, with the classifier trained on one set of annotated examples, and then tested on another set selected by random sampling. When this testing is done a single time, textbook techniques of point estimation, confidence interval estimation, and hypothesis testing on effectiveness are directly applicable.

In many practical circumstances, however, effectiveness is tested multiple times. The developer may repeatedly add examples to the training data, retrain classifiers, and test until estimated effectiveness reaches a target level. Or the classifier may already be constructed, and the developer wishes to validate its effectiveness while minimizing the number of test examples that need to be labeled. Or the developer may grow both training and tests sets at the same time, seeking both to improve and to certify effectiveness.

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

Unfortunately, as we observe in this paper, repeated testing of a classifier's effectiveness introduces statistical bias into estimates. Statistical inference is based on the recognition that any particular measurement has an element of randomness to it. We may by chance draw a sample that gives an optimistic estimate, leading us to put into use a classifier less good than it appears. Conversely, we may reject a classifier incorrectly based on a sample that gives too pessimistic an estimate. But by drawing a sample randomly we can quantify these risks, and reduce them through larger samples.

When we test repeatedly, however, the two dangers are not symmetric. An optimistic estimate will lead us to put a classifier into use and congratulate ourselves on having annotated little data. A pessimistic result will lead us to do more training, tuning and/or testing. We are therefore more likely to accept overestimates than underestimates in sequential testing, leading to an optimistic bias.

In this paper, we explore the behavior of sequential testing for the three scenarios outlined above (fixed test and variable training; fixed training and variable test; and both test and training variable). Two estimates of effectiveness are considered: a point estimate of the F_1 score, and the lower bound of a one-sided 95% confidence interval on F_1. We empirically quantify the bias introduced by sequential testing across a range of text classification tasks, provide guidance for practitioners, and point to a range of open questions.

2. MATERIALS AND METHODS

We use SVM^{perf} (V. 3.0) with a linear kernel [Joachims, 2006] to train text classifiers optimized for F_1, using flags "-c 1000 -l 1 -w 3". Our data set is the RCV1-v2 collection of Reuters newswire stories [Lewis et al., 2004]. Feature vectors were constructed using the RCV1-v2 stopped, stemmed token files (On-Line Appendix 12 of Lewis et al. [2004]). Feature values were log TF x IDF weights, with IDF equal to the natural log of the number of documents in the collection divided by the number of documents the word occurs in.

We defined 29 binary classification tasks corresponding to the 29 Topics categories with at least 25,000 positive examples in the 804,414-document collection. The use of high frequency categories enabled studying several orders of magnitude of variation in sample sizes. Our effectiveness measure is F_1; that is, the harmonic mean of precision and recall. We define F_1 to equal 1.0 if a classifier makes no positive predictions on a data set with no positive examples [Lewis, 1995].

Our experiments compare F_1 estimates with true population F_1. We derive a highly accurate estimate of the latter from a randomly sampled "truth set" of 700,000 of the RCV1-v2 documents. The mean width of the two-sided confidence interval on that estimate is 0.0059—minuscule compared to the variations we see in estimates from our smaller experimental test sets.

The test and training sets in our experiments are drawn from these 700,000 documents and from the remaining 104,414 documents respectively. In each run, the training and/or test set was grown in a nested fashion, by adding 20 randomly selected annotated documents, retraining the classifier on the training set if necessary, and estimating F_1 based on the test set. The 20 documents were either all added to test (Section 3.1), all added to training (Section 3.2), or half added to training and half to test (Section 3.3).

We compute two estimates of F_1 from each test set. First, the F_1 score on the test set is taken as a point estimate, $\hat{F}_1$, of the true F_1. Second, we find the lower bound, θ, of a lower one-sided 95% confidence interval on F_1. We compute this confidence interval estimate by taking the 5th percentile of 40,000 Monte Carlo draws on beta-binomial posteriors on the classifier's true positive and false positive rates, using the method described by Webber [2013] (see Goutte and Gaussier [2005] for a similar binomial posterior on F_1). (Similar sequential biases would arise for other interval methods, or other effectiveness measures such as precision or recall.)

Webber [2013] shows that this Monte Carlo procedure gives coverage probabilities very close to the nominal 95% level for two-tailed intervals on recall, in contrast to commonly used normal approximations. Nevertheless, we checked the accuracy of the method for one-tailed intervals on F_1. We calculated an *observed coverage*; namely, the proportion of testing points at which the lower bound of the confidence interval is at or below the true F_1 score. An edge case occurs when the classifier achieves no true positives on the test set, either because the classifier is poor or the test set has no positive examples. In this case, the lower bound of the confidence interval is 0.0, guaranteeing 100% coverage. To avoid crediting the algorithm for this trivial case, we computed coverage only on cases with at least 100 training and 100 test examples. With this restriction, we find that the observed coverage of Webber's algorithm is very near 95% in all experiments.

As will be seen, however, applying sequential stopping rules to nominal 95% confidence intervals leads to stopping conditions with actual coverage lower than 95% in all three scenarios studied. One way to adjust for this might be to compute the confidence interval at a nominal level higher than 95%, but assert only a 95% level for the resulting interval. For each protocol, we calculate what the nominal level must be to achieve 95% coverage with our data set.

3. RESULTS

We present results for three conditions: variable test, variable training, and variable both test and training.

3.1 Variable Test, Fixed Training

First, we consider the case where a classifier has been produced using a fixed training set. A developer wishes to certify, at a given confidence level, that the classifier exceeds a target value of effectiveness. We assume that a testing budget of 10,000 annotations is available, but that the developer would prefer to pay for fewer annotations if possible. This is the standard setting in sequential sampling theory [Wetherill and Glazebrook, 1986]. We assume that they iteratively annotate randomly selected examples, add them to the test data, and compute the two estimates of F_1 (point estimate and lower bound of a one-sided lower 95% confidence interval). The developer decides to accept the classifier (and stop testing) if at some point the lower confidence limit exceeds the target value of F_1. They reason (erroneously) that they will accept an inadequate classifier at most 5% of the time. They reject the classifier if the budget of 10,000 annotations is used up without accepting.

Figure 1 shows the results of one such classifier evaluation run. The target value of F_1 is 0.5, while the true effectiveness (unbeknownst to the developer) is 0.42. The point estimate $\hat{F}_1$ happens to almost always be above its true value, and for many test set sizes above the substantially higher target value. The lower bound of the confidence interval is more conservative, but wanders above the true value at several points, and above the developer's target for several test sets in the range of 100 to 200 documents. The developer would have mistakenly halted testing and accepted the classifier upon reaching the test set of size 120. For large test sets, both the point estimate and the lower bound of the confidence interval approach the true value of effectiveness, but our developer would never reach these test set sizes. Figure 1 is only for a single run, and perhaps (from a reliable evaluation perspective) an unlucky one. A 95% confidence interval is expected to fail roughly 5% of the time. We therefore performed the same experiment across all 29 topics, and made 20 runs for each topic. In each case the target effectiveness was set infinitesimally above the true effectiveness of the classifier, so that the classifier should be rejected in every run; not doing so counts as a failed evaluation. Unsurprisingly, the point estimate exceeded the target value at some point on almost every run. So using the point estimate would fail almost 100% of the time. But even the lower bound of the confidence interval exceeded the target value 31.55% of the time (far more than the nominal 5% the developer might expect), causing the classifier to be erroneously accepted. The observed coverage of the confidence intervals was 95.68%, making clear that the problem is the developer's sequential stopping rule, not the (single-test) confidence interval estimate.

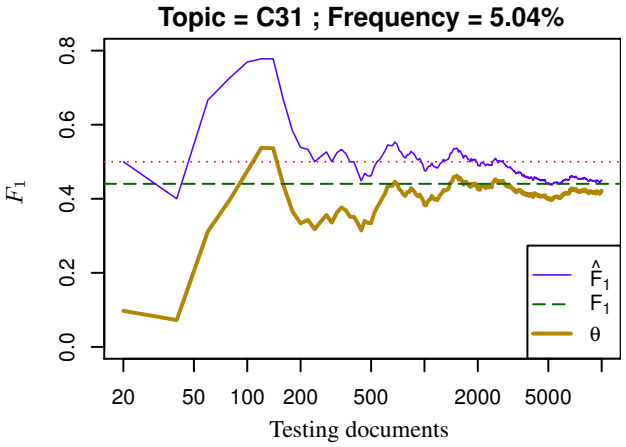

Figure 1: Variable test, fixed training. The dashed line is true effectiveness (F_1); the dotted line is the target F_1; the solid line is estimated F_1, for increasing test set sizes; and the bold line is the lower bound of the 95% one-sided confidence interval on F_1. Results are shown for a single run for Topic C31.

While the bias introduced by overfitting has long been a subject of study in machine learning [Toussaint, 1974], the bias introduced by sequential testing appears to have been ignored. Sequential stopping rules, and techniques for unbiasing estimates when using them, have been studied in statistics and quality control [Siegmund, 1985, Wetherill and Glazebrook, 1986]. The theory is well developed (if not simple) for cases such as the binomial proportion and the normal mean. Unfortunately, F_1, as the ratio of two unknown quantities, is substantially more difficult to address analytically. We can, however, observe an empirical adjustment on our dataset. To achieve actual 95% confidence, assuming that lower-bound effectiveness is tested every 20 documents, we should set a nominal confidence level of 99.5%.

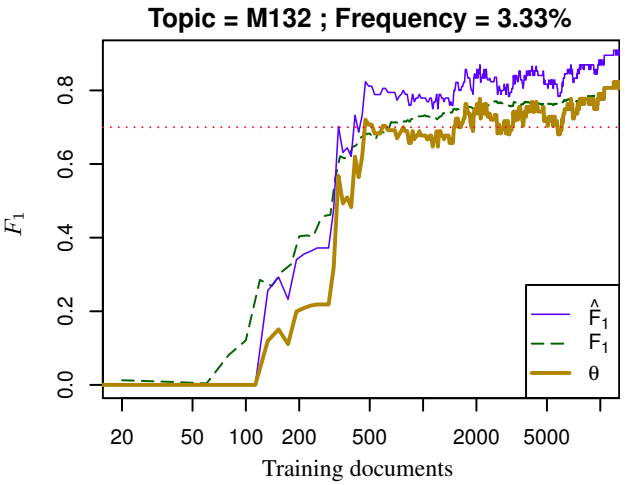

Figure 2: Fixed test, variable training. True classifier effectiveness (measured by F_1) as the training set is increased is shown as a dashed green line; the estimated effectiveness, based on a test set of $1{,}067$, is the solid blue line; and the 95% lower-bound confidence on that effectiveness is the bold gold line.

3.2 Fixed Test, Variable Training

Next we examine a common use of sequential testing in classifier construction. The developer of the classifier has a fixed, randomly selected test set. Training examples are selected, annotated, and used to improve the classifier. At regular intervals, the classifier is tested on the test set. We assume that training and testing continue until the estimated effectiveness of the classifier reaches some target value, or when a total of $15{,}000$ documents have been labeled. We fix the test set size at $1{,}067$ documents (a commonly used value that gives a maximum two-sided confidence interval width of $\pm 3\%$ on a binomial proportion). This corresponds to training budgets between 13 and 13,933, with increments of 20 documents. However, the choice of a particular test set size is not critical to our results.

Figure 2 shows an example of this scenario. The developer has set a target value of 0.7 on F_1. Thinking themselves conservative, they decide to stop training when the lower confidence limit on F_1 exceeds 0.7. They (erroneously) reason they will accept an inadequate classifier at most 5% of the time they use this procedure. They reject the classifier if the combined budget of $15{,}000$ annotations is used up without accepting.

The true effectiveness of the classifier rises fairly smoothly with the addition of training data, as we would expect. The two estimates of effectiveness are, however, far more erratic. Both the point estimate (at 333 training examples) and, surprisingly, the lower bound of the confidence interval (at 473 examples) exceed the target value well before the true effectiveness does (at 660 examples).

The erratic behavior evinced in Figure 2 seems at first surprising given a fixed test set and steady growth in true effectiveness. However, while the true effectiveness of successive classifiers increases fairly steadily as the training set grows in size, their behavior on a particular test set of modest size can fluctuate greatly from classifier to classifier. It is not the case that training improvements smoothly and monotonically convert test set errors into test set successes. Thus, even though no additional sampling is being performed, random fluctuations in estimated effectiveness occur, and we are more likely to stop when those fluctuations lead to overstating effectiveness than when they lead to understating it.

Does Figure 2 simply show an unlucky case? We again ran 29 topics and 20 runs per topic. In each run, we set the target value for effectiveness to be 90% of the highest true effectiveness achieved (usually at the largest training set size) during the run. The effectiveness of the classifier was estimated after each addition of 20 annotated examples to the training set. We examined two stopping rules: stopping when the point estimate first hits this target, and stopping when the lower confidence interval bound first hits this target. For unbiased point estimates, we might expect the first rule to stop early 50% of the time (the point estimate is as likely to be above as below the true effectiveness), and the latter rule to stop early only 5% of the time. In practice, across 29 topics and 20 runs per topic, we found that the point-estimate rule stopped early 53.58% of the time. The lower-bound stopping rule stops early 8.13% of the time. The latter of these values is significantly different from the expected proportion ($p < 0.01$ under an exact binomial test with $29 * 20 = 580$ observations); the former is not significantly different, but the jaggedness of the estimate curve means that there inevitably is an optimistic bias, though our number of categories and runs was too small to demonstrate this at the $p = 0.05$ level. The observed confidence interval coverage for this scenario is 95.26%, so again it is the stopping rules that are the problem.

This bias is smaller than in Section 3.1, but also difficult to attack analytically. The theory of sequential testing has focused on estimating fixed distributions, not time varying ones. Computational learning theory has found the analysis of learning curves challenging even in the generalization framework (corresponding to our comparatively smooth curve of true effectiveness) [Haussler et al., 1996]—we do not know of a treatment of learning curves for finite test sets. Lacking analytical guidance, we observe that a nominal confidence level of 97% is required to achieve an actual confidence of 95%.

3.3 Variable Test, Variable Training

Finally, we consider the situation in which both test and training set are increased over time. Though not a widely used approach, it has the advantage of avoiding committing to too small or too large a testing or training set size at the outset. Many adaptive policies are possible; we consider the simple one of allocating as many documents to testing as we do to training (for our experiments, 10 to training and 10 to test at each increment, then performing the test).

Figure 3 shows what happens for a single run on a single topic under this regime. We assume the developer has set a target value of 0.6 on F_1, and decides to stop training and testing when the lower confidence limit exceeds this target. They reject the classifier if the combined budget of $15{,}000$ annotations is used up first.

Again, the increase in true effectiveness with increasing training set sizes is relatively smooth. And again, the point estimate on effectiveness is more erratic, here from a combination of both changing classifiers (as the training set is increased) and the changing content of the test set. The lower bound once more varies with the point estimate, shrinking the gap gradually, since half of the annotations go to testing. Both the point estimate of F_1 and the lower bound of the confidence interval hit the target of 0.6 before true effectiveness reaches it: the developer would stop too early.

We once more estimate the degree of bias by observing behavior across all 29 topics and 20 runs. Setting the target effectiveness at 90% of maximum actual effectiveness, we find that the point estimate crosses this threshold prematurely 68.38% of the time, while the lower bound crosses it prematurely 9.40% of the time. The observed coverage for this scenario is 95.39%, showing again that stopping rules are at fault. Here, a nominal confidence level of 98% is required to achieve an actual confidence of 95%.

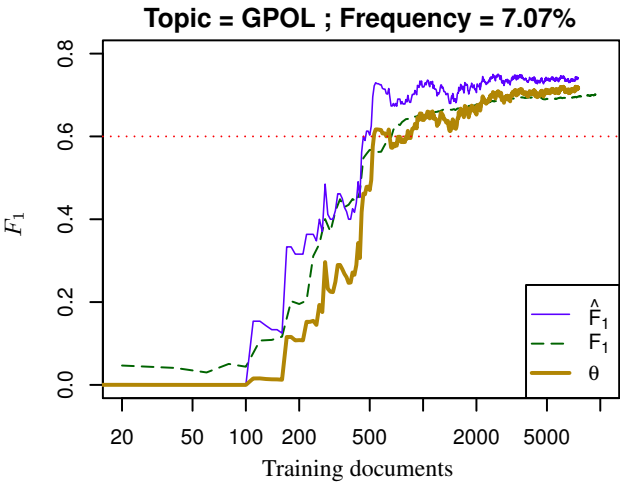

Figure 3: Variable test, variable training. The dashed green line gives true effectiveness (F_1) of the classifier at each training set size. The solid blue line gives a point estimate on effectiveness of a test set as large as the training set at each point. The bold gold line is the lower bound of a 95% lower one-sided confidence interval estimate.

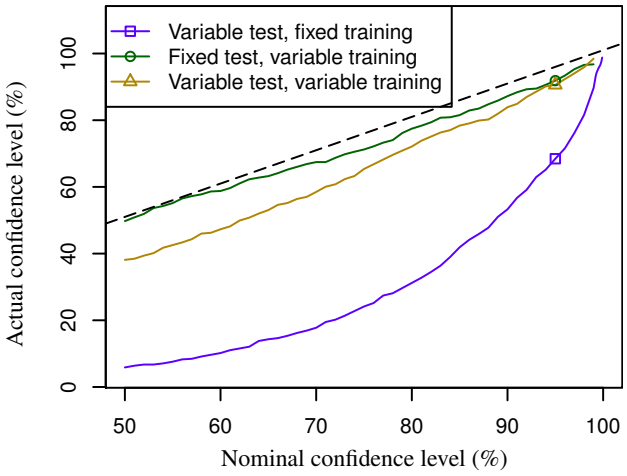

Figure 4: Actual confidence corresponding to nominal confidence level.

3.4 Empirical Bias

To generalize our observations on the nominal confidence of 95%, we tried a range of confidence levels between 50% and 99%. Figure 4 shows the actual confidence of the three scenarios studied in this paper. It is, in all cases, lower than the nominal confidence (as shown by the dashed line), and most markedly so for the variable test, fixed training case. These curves can serve as an indicator to correct the bias resulting from sequential sampling, though it is unclear how well they generalize to other conditions.

4. CONCLUSIONS AND FUTURE WORK

We have examined the bias involved in three sequential testing methods: fixed training and variable test; fixed test and variable training; and both test and training variable. The bias in fixed training and variable test is severe: a nominal 95% confidence interval provides only 68% coverage in practice, for the scenario considered here. Variable training and fixed test also introduces a significant bias to estimates, with a nominal 95% giving 92% coverage. The combination of variable training and variable test has only slightly greater bias than variable training with fixed test; a nominal coverage interval of 95% gives actual coverage of 91%. These are not due to problems in the Monte Carlo method of calculating intervals on F_1, which gives coverage very close to nominal levels for all three conditions. The bias lies solely in the use of sequential testing.

For the case of fixed training (and fixed classifiers in general) and variable test a sequential sampling analysis of the F-measure would be of great practical value, enabling valid estimates while saving labeling costs. Both traditional analytical approaches (e.g. Brownian motion approximations [Siegmund, 1985]) and simulation techniques are worth considering. For the cases in which the training set varies, estimation must track the moving target of the learning curve. Analytical solutions are likely impossible, and even simulation techniques will require confronting issues of classifier stability [Kearns and Ron, 1999].

In this paper, we have provided empirical adjustments to confidence limits, but these apply only to the particular scenarios and dataset explored here. For other cases, they may provide some rule-of-thumb guidance for developers, but they cannot stand in place of a statistically well grounded validation. For now, a trustworthy validation requires testing the final classifier once on a set of annotated instances that is used for no other purpose.

Acknowledgments

This material is based 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

Cyril Goutte and Eric Gaussier. A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. In *ECIR '05*, pages 345–359, March 2005.

David Haussler, Michael Kearns, H Sebastian Seung, and Naftali Tishby. Rigorous learning curve bounds from statistical mechanics. *Machine Learning*, 25(2):195–236, 1996.

Thorsten Joachims. Training linear SVMs in linear time. In *ACM KDD '06*, pages 217–226, 2006. ISBN 1-59593-339-5.

Michael Kearns and Dana Ron. Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. *Neural Computation*, 11(6): 1427–1453, 1999.

David D Lewis. Evaluating and optimizing autonomous text classification systems. In *ACM SIGIR '95*, pages 246–254, 1995.

David D. Lewis, Yiming Yang, Tony G. Rose, and Fan Li. RCV1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5:361–397, December 2004.

David Siegmund. *Sequential Analysis: Tests and Confidence Intervals*. Springer-Verlag, 1985.

Godfried T. Toussaint. Bibliography on estimation of misclassification. *IEEE Transactions on Information Theory*, IT-20(4):472–479, July 1974.

William Webber. Approximate recall confidence intervals. *ACM TOIS*, 31 (1):2:1–2:33, January 2013.

G. Barrie Wetherill and Kevin D. Glazebrook. *Sequential Methods in Statistics*. Chapman and Hall, 3rd edition, 1986.