Improving Active Learning Recall via Disjunctive Boolean Constraints

Emre Velipasaoglu
Yahoo! Inc.
Sunnyvale, CA 94089
USA
emrev@yahoo-inc.com

Hinrich Schütze
Institute for NLP
University of Stuttgart
Germany
hs999@ifnlp.org

Jan O. Pedersen
Yahoo! Inc.
Sunnyvale, CA 94089
USA
jpederse@yahoo-inc.com

ABSTRACT
Active learning efficiently hones in on the decision boundary between relevant and irrelevant documents, but in the process can miss entire clusters of relevant documents, yielding classifiers with low recall. In this paper, we propose a method to increase active learning recall by constraining sampling to a document subset rich in relevant examples.

Categories and Subject Descriptors: I.2.6 [Artificial Intelligence]: Learning; H.3.1 [Information Storage and Retrieval]: Content Analysis and Indexing.

General Terms: Algorithms, Performance.

Keywords: Practical text classification, active learning, missed cluster effect.

1. INTRODUCTION
Machine learned text classifiers are commonly created by estimating the parameters of a statistical classification model on a labeled training set [7]. Active learning (AL) is an efficient training set creation method. AL starts with a seed set, iteratively samples unlabeled examples from the subspace that contains "informative" or "uncertain" documents, labels and adds them to the training set, and retrains the classifier thus redefining the region of uncertainty in each step [1]. It has been shown that AL is effective in creating high performance classifiers with a relatively small number of labeling decisions [2].

Unfortunately, AL can produce classifiers low in recall because the sampling scheme hones in very rapidly on the decision boundary of a classifier produced in an early iteration. If there is only one cluster of relevant documents, this procedure is likely to rapidly discriminate between relevant and non-relevant documents. However, if there are several discrete clusters of relevant documents it is possible that whole clusters will be missed, ruining recall if not precision. We refer to this phenomenon as the *missed cluster effect*. We have found that about a third of relevant documents in the Reuters corpus are in missed clusters after 100 iterations of AL [6].

This paper addresses the missed clusters using a bootstrap set of documents that can be obtained by AL. We propose a method to constrain the document space AL samples from in order to increase the population of relevant examples. This modification improves the likelihood that AL will sample from otherwise missed clusters, resulting in significant improvement to recall.

SIGIR'07, July 23–27, 2007, Amsterdam, The Netherlands.
ACM 978-1-59593-597-7/07/0007.

2. METHOD
The basic idea of our method is to automatically find a set of *characteristic terms* for recall enhancement by analyzing the labeled set after a number of iterations of AL. Terms that occur more frequently in the labeled set than in the pool are particularly interesting. These terms represent concepts that may co-occur with other terms to yield concepts relevant to missed clusters. We propose to use a disjunctive query of such terms to obtain a sample of unlabeled documents from the pool. This subset is empirically found to be richer in missed cluster documents than the unlabeled pool. We modify AL simply by constraining the candidate documents to this subset. A new set of characteristic terms is estimated for each subsequent iteration, and a disjunctive query is issued to define a new subset of unlabeled documents.

We use a 2-stage method to estimate the set of characteristic terms based on the counts for each term in Table 1 and algorithm below.

1. Order terms t_i by descending $\chi^2_{t_i} = \chi^2_{t_i}\left(A_{t_i}, B_{t_i}, C_{t_i}, D_{t_i}\right)$
2. Select first m terms: $T_{stage-1} = \{t_i | i \le m\}$
3. Order terms t_i in $T_{stage-1}$ by ascending $r_{t_i} = \left|1 - \left(A_{t_i} / C_{t_i}\right)\right|$
4. Select first n terms: $T_{stage-2} = \{t_i | i \le n\}$

In words, we select top m terms by descending chi-square statistic in stage-1, and top n terms within that set using the criterion that ratio of counts A and C is as close to 1 as possible in stage-2.

Table 1. Counts used to estimate characteristic terms:

A_{t_i}	# docs in **labeled** portion of pool **containing** term t_i.
B_{t_i}	# docs in **unlabeled** portion of pool **containing** term t_i.
C_{t_i}	# docs in **labeled** portion of pool **not containing** term t_i.
D_{t_i}	# docs in **unlabeled** portion of pool **not containing** term t_i.

Analysis of a case illustrates how the method works. For example, the GVIO category trained on all labeled examples has recall of 77%, whereas ordinary AL (i.e. uncertainty sampling as in [4]) only reached 54% in 200 iterations. Recall deficiency signals the possibility of missed clusters. At the 50th iteration, the rate of relevant examples in the subset of unlabeled documents selected by the disjunctive query of the characteristic terms is about 15 times that of the unlabeled pool. Further analysis shows that most of these relevant documents are false negatives. As a result AL has higher likelihood to sample the missed clusters when constrained to this subset of documents.

3. EXPERIMENTS

The first half of the RCV1 corpus (400,001 documents) was randomly split into POOL and EVAL. Seed and query documents were drawn from POOL, and EVAL was used for evaluation. 43 categories with frequencies ranging from 0.004% to 47% were selected, representing a mix of "small" and "large" categories. Standard tf-idf vector space representation was used. Details of this experimental setup can be found in [6]. Linear SVMs were used as they are viewed as having close to optimal performance in text categorization [8]. We used uncertainty sampling as in [4].

For each category, we ran 200 iterations of AL starting with a seed set of 5 positive and 5 negative documents selected randomly from POOL. The first 50 iterations were ordinary AL (uncertainty sampling). After that, we constrained AL as explained above. In estimating the characteristic terms, we selected top m=100 terms in stage-1 and n=10 terms in the stage-2.

Table 2. Performances of ordinary and constrained AL.

	Ordinary AL				Constrained AL (% Relative Gain)			
Category	**+**	**F**	**P**	**R**	**+**	**F**	**P**	**R**
SEASIA	5	0	0	0	20	0	0	0
I22470	6	0	0	0	0	0	0	0
I64600	7	0	0	0	0	0	0	0
ISLAM	7	0	0	0	0	0	0	0
SPSAH	10	93	100	88	0	0	0	0
I6540030	12	43	86	29	0	0	0	0
I25520	13	38	100	23	0	0	0	0
SURM	13	70	100	54	0	0	0	0
I22300	15	26	100	15	0	0	0	0
I32550	15	20	100	11	0	0	0	0
I36102	15	12	100	6	0	0	0	0
I45000	17	31	63	21	0	0	0	0
GABON	20	73	100	57	0	0	0	0
I97412	22	24	94	13	0	0	0	0
ERTRA	28	68	93	53	0	0	0	0
I45300	30	9	87	5	7	-4	-1	-4
NEPAL	30	94	99	90	0	0	0	0
SENEG	34	67	89	53	3	2	0	3
I8500031	37	26	69	16	14	35	14	41
PARA	41	77	92	66	0	-1	0	-2
BOL	43	70	84	60	-7	-3	3	-7
I83940	48	20	84	11	4	1	0	1
I82002	49	22	93	12	6	11	-4	14
E411	56	63	72	55	5	2	2	2
SLVNIA	56	89	98	81	-5	-4	1	-7
I81402	57	4	67	2	21	9	11	9
I42700	63	55	85	41	5	3	-3	7
I13000	64	37	86	24	30	27	-5	40
I42900	64	60	82	47	5	-5	4	-10
I65600	64	32	85	20	11	11	-10	18
M131	64	76	93	65	9	4	-6	11
M132	65	77	87	70	11	1	2	0
EEC	66	76	92	65	24	6	-2	13
GVIO	66	65	83	54	38	11	-8	28
DEN	67	88	97	81	9	1	0	1
E51	67	36	84	23	25	26	-12	43
ROM	68	93	99	87	21	0	0	1
E212	69	77	92	66	9	1	-2	3
CANA	81	78	95	67	4	5	-3	11
AUSTR	83	88	97	81	13	2	-2	6
GCAT	101	88	90	85	-1	0	2	-3
CCAT	105	84	83	86	-1	-1	-1	-1
USA	106	84	93	77	-2	-2	-7	3

4. RESULTS

The number of relevant examples (+) at the end of 200 iterations, precision (P), recall (R) and F measured on EVAL are compared for ordinary and constrained AL. In Table 2, absolute performance is listed for ordinary AL and relative gain, calculated as (Y-X)/X, listed for constrained AL. For instance, the E51 category had 67 relevant examples at the end of 200 iterations of ordinary AL (including 5 relevant examples in the seed set). The performance numbers were 36%, 84% and 23% for F, P and R, respectively. Constrained AL produced 25% more relevant examples, increased F and R by 26% and 43%, respectively, and caused a 12% drop in P. Figure 1 shows the relative gain in F vs. the final number of relevant examples in ordinary AL.

Average relative gain in recall is 5.1% (p-value = 0.008 in one-sided paired t-test). F is up 3.2% (p-value = 0.008) and precision is down 0.6% but not significant (p-value = 0.1). There was no difference in performance between ordinary and constrained AL for small categories. Constrained AL caused a slight loss of accuracy for the largest 3 categories, GCAT, CCAT and USA, which have population rates 30%, 47% and 33%, respectively. It appears constraining by a disjunctive query of only 10 terms is too aggressive and unnecessary for these categories. A strong gain is observed for "medium" size categories, where the population rate ranged from 0.04% to 4.3%. In general, improvement in recall came at the expense of a small loss in precision. This is acceptable and even preferred in many practical text classification scenarios, since ordinary AL is heavily biased for high precision [6].

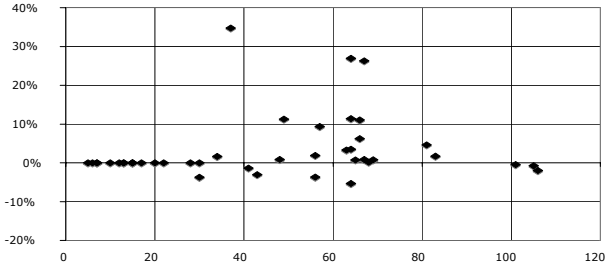

Figure 1. Number of relevant examples in ordinary AL vs. percent relative gain in F by constrained AL.

5. RELATED WORK

In [3], a prior distribution over terms is utilized to increase recall. In [5], AL is extended by requesting explicit feedback about terms in interleaved steps. Unlike ours, these solutions require explicit information over term space. Our method would be useful where term space knowledge is not clear and can be complementary to methods that leverage domain knowledge explicitly.

6. REFERENCES

[1] D. Cohn, L. Atlas, and R. Ladner. Improving generalization with active learning. Mach. Learn., 15(2):201–221, 1994.

[2] S. Dasgupta. Analysis of a greedy active learning strategy. NIPS, 2004.

[3] A. A. Dayanik, D. D. Lewis, D. Madigan, V. Menkov, A. Genkin. Constructing informative prior distributions from domain knowledge in text classification. SIGIR, 2006.

[4] D. D. Lewis and W. A. Gale. A sequential algorithm for training text classifiers. SIGIR, 1994.

[5] H. Raghavan, O. Madani, R. Jones. InterActive Feature Selection. IJCAI, 2005.

[6] H. Schütze, E. Velipasaoglu, and J. Pedersen. Performance thresholding in practical text classification. CIKM, 2006.

[7] F. Sebastiani. Machine learning in automated text categorization. ACM Comp. Surveys, 34(1):1–47, 2002.

[8] Y. Yang and X. Liu. A reexamination of text categorization methods. SIGIR, 1999