Batch Active Learning for Text Classification and Sentiment Analysis

Anonymous ACL submission

Abstract

Supervised learning of classifier for text classification and sentiment analysis relies on availability of labels that may be difficult/expensive to obtain. Active learning techniques optimize the process of obtaining labels by sequentially selecting documents from unlabels set for which the labels would be most valuable. Batch active leaning selects a batch of documents for labeling which is much more demanding. In this paper we propose a new methods for batch active learning by combining Bayesian strategies with agglomerative clustering. We study the effect of this proposition in large scale study comparing the effect of varying distinct factors of active learning algorithm (initialization of the algorithm, uncertainty representation, acquisition function and batch selection strategy). Various combinations of these are tested on selected NLP problem with documents encoded using RoBERTa. Datasets cover context integrity, fake news detection and sentiment classification. We show that each of the active leaning factor has advantages for certain datasets or experimental setting.

1 Introduction

Supervised learning of classifiers relies on availability of class labels which often involves human annotator for majority of NLP tasks. This can be costly for large datasets. Active learning is a strategy designed to minimize this cost by automatic selection of those unlabeled documents that expected to bring useful information for the classifier. Advantages of this approach have been demonstrated even for classical methods such as SVM (Tong and Koller, 2001). The most conventional active learning strategies select only one unlabeled document after each training round to query due to simplicity of its selection. The next query document is selected only after the first on is labeled and the model re-trained which means that the annotator has to wait for retraining. This impractical strategy can be avoided is the active learning algorithm selects a batch of documents. Novel methods for *batch active learning* appear frequently, each demonstrating advantages on their benchmark data.

One of the recent approaches demonstrate effectiveness even for batches of 5000 samples (Citovsky et al., 2021), combining the min-margin acquisition function with clustering under name Hierarchical Agglomerative Clustering (HAC). In this contribution, we propose novel modifications of this idea using alternative acquisition functions and investigate their performance. Specifically we propose to extend the HAC approach to Bayesian setting by replacing the min-margin by Bayesian acquisition function, BALD (Houlsby et al., 2011). However, the size of the minibatch is only one of many factors in performace of the active learning algorithms. Other factors are: i) selected acquisition function, ii) representation of uncertainty of the classifier, and iii) initialization of the network. The key contribution of our work is sensitivity study of the classification task to these factors over a range of datasets from various text classification tasks.

Various comparative studies have been performed recently with various focus and results. Batch active learning was studied in (Dor et al., 2020) for only one size of the batch (50 documents per query). Large sensitivity to the type of dataset was reported in (Prabhu et al., 2021), where different method won for different data. Large variability of the results was also observed in (Jacobs et al., 2021). In (Schröder et al., 2021), the min-margin strategy was shown to be competitive to prediction entropy based method on a range of embeddings. The comparative studies shared similar properties, such as fixed network for embeddings (improvement with re-training can be expected (Margatina et al., 2021) but maybe too costly). All studies also assume cold start, i.e. completely new initialization of the classifier after each round of querying. This is motivated by the fear of over fitting which was demonstrated in (Hu et al., 2018) for hot start, i.e. continuation of training of the classifier. A compromise in the form of warm-start, i.e. adding noise to the weights of the previous classifier, was proposed in (Sahan et al., 2021).

The paper is organized as follows. In Section 2, we briefly review all tested factors of the batch active learning. Experimental setup of the sensitivity study is described in Section 3 and results are reported in Section 4.

Algorithm 1 General batch active learning

Initialize: set classifier structure $\mathbf{y} = \mathbf{y}(\mathbf{x}, \theta)$, iteration counter $i = 0$, initial data $\mathbf{Y}^{(0)}, \mathbf{X}^{(0)}, \mathbf{X}_u^{(0)}$
Iterate until a stopping condition:
1. train a classifier parameter $\theta^{(i)}$ on $\mathbf{Y}^{(i)}, \mathbf{X}^{(i)}$, starting from $\theta_{\text{init}}^{(i)}$
2. compute the value of a label for all documents in the unlabeled dataset, $a_l = A(\mathbf{x}_l, \theta^{(i)}), \forall \mathbf{x}_l \in \mathbf{X}_u^{(i)}$
3. select a batch of documents, $\tilde{\mathbf{X}} \subset \mathbf{X}$, for labeling using a_l
4. obtain $\tilde{\mathbf{y}}$ for $\tilde{\mathbf{X}}$ and extend the training set $\mathbf{X}^{(i+1)} = \mathbf{X}^{(i)} \cup \tilde{\mathbf{X}}, \mathbf{y}^{(i+1)} = \mathbf{y}^{(i)} \cup \tilde{\mathbf{y}}, i = i + 1$.

2 Batch Active Learning Methods

Throughout the paper, we will use the RoBERTa embedding (Liu et al., 2019) to represent documents in the feature space. RoBERTa is a modified BERT transformer model (Devlin et al., 2018) that achieved comparable performance to BERT in (Schröder et al., 2021) and outperformed all other embedding in (Sahan et al., 2021). Representation of the k-th text document $\mathbf{x}_k$ is calculated as the mean value from sentence embeddings of all sentences in the text.

The aim of document classification is to find a classifier $\hat{\mathbf{y}} = \mathbf{y}(\theta, \mathbf{x})$ predicting the class label for each document representation $\mathbf{x}$. In supervised setting, the classifier parameters are found on a training set $\{\mathbf{x}_k, \mathbf{y}_k\}_{k=1}^K$ by matching the prediction $\mathbf{y}(\theta, \mathbf{x}_k)$ with the provided label $\mathbf{y}_k$ for each document. We are concerned with binary classification for simplicity, however, an extension to multiclass is straightforward.

We assume that for the full corpus of text documents $\mathbf{X}$, only a small initial set of labels $\mathbf{Y}^{(0)}$, is available. The full set $\mathbf{X}$ is thus split to the labeled, $\mathbf{X}^{(0)}$, and unlabeled parts, $\mathbf{X}_u^{(0)} = \mathbf{X} \backslash \mathbf{X}^{(0)}$, the training set in the first round is then $\mathbf{X}^{(0)}, \mathbf{Y}^{(0)}$. Active learning is defined as a sequential extension of the training data set following a simple iterative strategy in algorithm 1.

The general algorithm can be specialized to many variants depending on various factor as specified next. We will introduce several choices labeled by the step in which they appear in algorithm 1.

1a. Uncertainty representation: The uncertainty can be represented by a maximum likelihood estimate, represented by a single network, or Bayesian probabilistic estimate, represented typically by an ensemble of networks. We will consider the following options: **Single network** with softmax output layer predicting normalized probability of each class in one hot encoding. This probability is conditioned on the parameter, and thus captures only aleatoric uncertainty. Uncertainty in parameters is not represneted. **Ensemble of networks**, represent uncertainty in parameters by different parameter value in each ensemble thus capturing both aleatoric and epistemic uncertainty. We consider two method for generating the ensemble members: i) *MC dropout* (Gal et al., 2017), where ensemble members are generated by random draws of of the dropout layers, and ii) *deep ensembles* (Lakshminarayanan et al., 2017) where ensembles are trained independently. Note that MC dropout is computationally much cheaper.

1b. Initialization of the training: Each training in step 1 is a new task. However, the data set typically overlaps with the one from the previous iteration, which motivates the following strategies of reusing results from the previous iteration. **Cold start** strategy is not reusing any information, the networks are initialized by random numbers, $\theta_{\text{init}}^{(i)} = \mathcal{N}(0, \sigma)$, where σ is given by the standard network init strategy, used most often (Dor et al., 2020; Citovsky et al., 2021; Schröder et al., 2021). **Hot start** strategy reuses all information, setting the estimate from the previous iteration as a starting point, $\theta_{\text{init}}^{(i)} = \theta^{(i-1)}$,criticized in (Hu et al., 2018). **Warm start** strategy a combination of the above, $\theta_{\text{init}}^{(i)} = \theta^{(i-1)} + \mathcal{N}(0, \sigma)$, where σ is a hyper-parameter (Sahan et al., 2021).

2. Acquisition function: Is a measure of the expected utility of the knowledge label, $\boldsymbol{y}_l$, for each document, $\boldsymbol{x}_l$, in the unlabeled data set. Different running index l is used to indicate that we operate on the unlabeled set. While many different utilities are proposed, we will study only the most popular ones. **Entropy** exists in two forms, entropy of the prediction $a_l = \mathbb{H}(y|\mathbf{x}_l, \theta)$ for a single network, or expected entropy $a_l = \mathbb{E}_\theta \mathbb{H}(y|\mathbf{x}_l, \theta)$ for ensembles. **BALD** is a mutual information metric, $a_l = \mathbb{E}_\theta \mathbb{H}(y|\mathbf{x}_l, \theta) - \mathbb{H}(y|\mathbf{x}_l)$ that is meaningful only for the ensembles. **Min-margin** is a minimum of difference between class predictions $a_l = -\min_{c,d \in [1,C]}(y_c - y_d)$ where C is the number of classes. Note carefully that extreme of this criteria is equivalent to maximum entropy for binary classification with single network.

3. Batch selection strategy When only one sample is to be selected it is optimal to choose the one with maximum utility given by the acquisition function. However, complexity of the maximum utility grows exponentially when the strategy has to select a batch of b documents for off-line labeling. Strategies that tries to approximate this selection using greedy search (Kirsch et al., 2019) are still too computationally expensive for large batches. Therefore, we select two batch selection strategies that scale well with b. **Top** selects top b samples from sorted values of a_l. This approach may select samples close to each other which are redundant. **HAC** is a strategy based on hierarchical clustering of a_l proposed in (Citovsky et al., 2021), and selecting top b samples from different clusters.

Tested algorithm variants: From the range of all possibilities, we will study the combinations that are existing in the literature: HAC min-margin using cold start (Citovsky et al., 2021), MC dropout with Entropy and BALD criteria using cold start (Gal et al.,

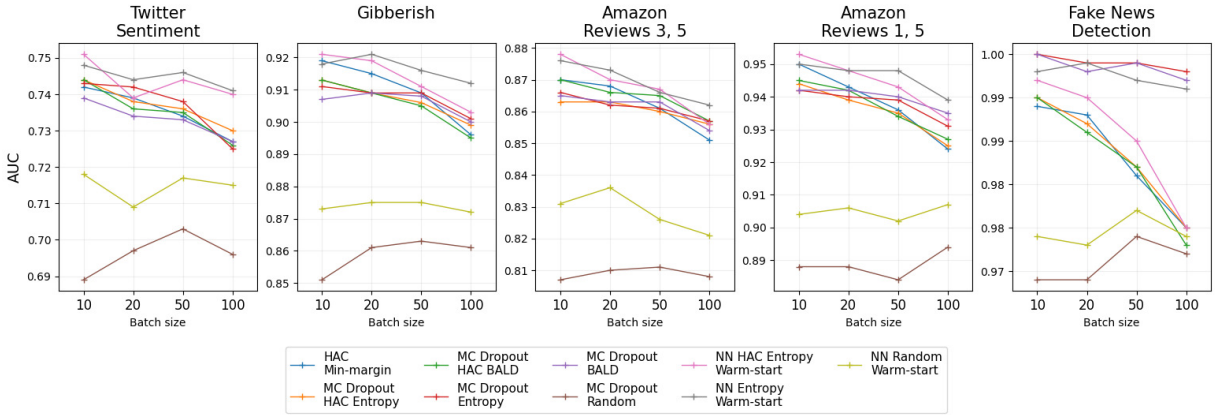

Figure 1: AUC metrics for 7 active learning and 2 random strategies after 1000 acquired samples given datasets and batch size.

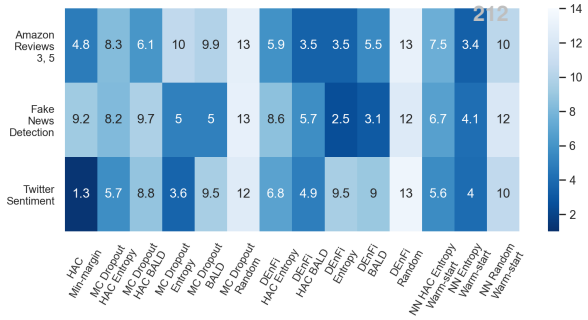

Figure 2: Aggregated mean rank for 14 tuples of learning algorithms and acquisition functions given Amazon Reviews 3,5, Fake News Detection and Twitter Sentiment for 50 active learning iterations with batch size 20.

2017), warm start ensemble learning with Entropy and BALD called DEnFi (Sahan et al., 2021), and conventional single-network with prediction Entropy with warm start. If HAC is not in the name, the Top strategy is used.

Since HAC strategy is an orthogonal factor to the remaining ones, we propose its combination with other approaches, giving rise to: HAC Entropy for the single neural network and both ensemble methods (MC dropout an DEnFi) and HAC BALD for the ensembles.

3 Experiment Setup

The methods were compared on different datasets and different batch size. The used datasets are positive/negative tweets from the Tweets [11] , Fake News Detection [8] , two pairs of categories from Amazon Reviews and Gibberish datasets. From all datasets, we select from 10000 text documents (5000 text documents per category, selecting only two categories for binary classification, e.g. 1 and 5 in Amazon reviews), which were the initial 10000 documents of the datasets given categories.

All algorithms were compared on area under the curve (AUC) on the test data (i.e. the documents not present in the training set). The algorithms were compared after 100 labels. The algorithms with smaller batch sizes thus benefited from higher number of retrainings. To reduce the influence of stochastic initialization and training, the AUCs were run 5 times and averaged. Even then, the difference between the algorithms were sometimes marginal. To show the effect of various factors on the performance, we sorted the AUC and assigned a rank of each method accordingly. I.e. the best performing method has rank 1, second rank 2, etc. This approach allows comparison of various methods across multiple datasets (Demšar, 2006) using order statistics. Intuitively: better method has lower average rank, methods with comparable ranks do not differ in performance.

4 Results

Parameter uncertainty: The effect of parameter uncertainty (Bayesian approach) is the most costly to evaluate, due to high computational demand of the ensemble approach (DEnFi). Therefore, we have evaluated all algorithm variant only for batch size $b = 20$. The results displayed in Figure 2. The advantage of the Bayesian approach are evident only for the Fake News dataset, in other datasets, DEnFi is not worth the computational cost and will be omitted from large scale studies. A summary of relative performance of all tested method for various batch sizes is displayed in Figure 3. Note that the datasets follow similar pattern, with the exception of the Fake News data sets, where the parametric uncertainty (now represented only be the MC dropout strategy) is beneficial.

Acquisition functions: Due to binary classification, the min-margin and entropy approaches coincide for single network function, which may also explain the results of (Schröder et al., 2021). The difference between Entropy and BALD for the ensemble methods seems insignificant.

Initialization of the training: The warm start strat-

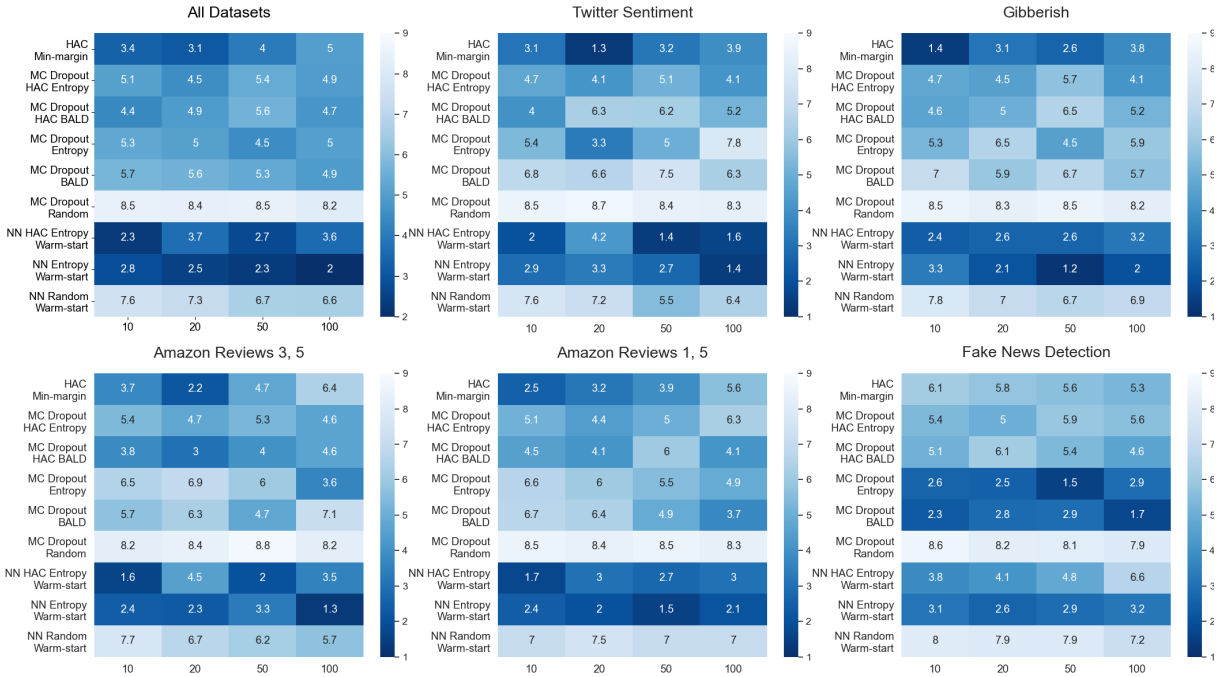

Figure 3: Aggregated rank for 7 active learning algorithms and two random strategies averaged over datasets as a function of different batch sizes.

egy (HAC Entropy warm start) is better or comparable in performance to cold start (HAC Min-margin); this is achieved at fraction of the training cost. This indicate that the additive noise is sufficient to avoid overfitting of the hot start (Hu et al., 2018).

Batch selection strategy: The HAC batch selection has clear advantage for lower batch sizes (10 and 20). This is consistent when comparing HAC and Top variants of all methods. This advantage diminishes for batch sizes of 50 and 100 where the top selection strategy achieves comparable (ensembles) of better (single NN) results. We conjecture that the most informative samples in our datasets are clustered in small clusters, hence selection of a batch with large enough size contains the relevant samples from sufficiently large areas.

5 Conclusion

We have studied influence of various factors of active learning algorithms on their performace on cover context integrity, fake news detection and sentiment classification tasks. While complex algorithms such as deep ensembles sometimes achieved good performance (Fake News detection), the winner on average is the classical prediction entropy of a single neural network with few proposed modifications. Specifically, the warm start of the network training achieves good performance at lower computational cost, and selection of the batch for annotation using agglomerative clustering improves performance for smaller batch sizes.

Appendix: Experiment parameters

Each active learning experiment was initialized by the training set $\mathbf{X}^{(0)}, \mathbf{Y}^{(0)}$ of 10 samples. The active learning strategy was set to sample b samples with discrete set of variants, $b = 10, 20, 50, 100$. The active learning was run until 1000 samples were labeled, i.e. making different number of step for each batch size (10 iteration for $b = 100$, 20 for $b = 50$, etc.). The batch selection follows the ϵ -greedy approach [25] , i.e. the samples selected by the acquisition function are accepted with probability $\epsilon = \frac{\exp(l-3)}{\exp(l-3)+1}$. A batch of random documents is selected for labeling if not accepted. After each request, the classification performance is evaluated on the remaining part of the selected dataset (i.e. on the 9990 text documents in the first evaluation) using the area under the ROC curve (AUC) metrics [9] . In order to make the results statistically valid, we repeat the described simulation loop 5 times for all datasets.

The initial number of epochs for the first iteration is set to 2500 for all algorithms. The same number is used for cold start strategy in each epoch. In the subsequent iterations, the weights of the previous estimate are perturbed by $\sigma = 0.3$ for MC dropout and $\sigma = 0.7$ for DEnFi. The training of the warm start strategies is run for 150 epochs. Both DEnFi and MC Dropout generate 5 ensemble members. The key difference is in computational complexity, while DEnFi has to do 150 epochs for each ensemble member, the MC dropout does it for only one network and generated ensemble members by 5 different realizations of the dropout mask.

References

Gui Citovsky, Giulia DeSalvo, Claudio Gentile, Lazaros Karydas, Anand Rajagopalan, Afshin Rostamizadeh, and Sanjiv Kumar. 2021. Batch active learning at scale. In *Advances in Neural Information Processing Systems*, volume 34, pages 11933–11944. Curran Associates, Inc.

Janez Demšar. 2006. Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research*, 7:1–30.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Liat Ein Dor, Alon Halfon, Ariel Gera, Eyal Shnarch, Lena Dankin, Leshem Choshen, Marina Danilevsky, Ranit Aharonov, Yoav Katz, and Noam Slonim. 2020. Active learning for bert: an empirical study. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7949–7962.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. 2017. Deep bayesian active learning with image data. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1183–1192. JMLR. org.

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. 2011. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*.

Peiyun Hu, Zachary C Lipton, Anima Anandkumar, and Deva Ramanan. 2018. Active learning with partial feedback. *arXiv preprint arXiv:1802.07427*.

Pieter Floris Jacobs, Gideon Maillette de Buy Wenniger, Marco Wiering, and Lambert Schomaker. 2021. Active learning for reducing labeling effort in text classification tasks. In *Benelux Conference on Artificial Intelligence*, pages 3–29. Springer.

Andreas Kirsch, Joost Van Amersfoort, and Yarin Gal. 2019. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. *Advances in neural information processing systems*, 32.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in neural information processing systems*, pages 6402–6413.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Katerina Margatina, Loic Barrault, and Nikolaos Aletras. 2021. Bayesian active learning with pretrained language models. *arXiv preprint arXiv:2104.08320*.

Sumanth Prabhu, Moosa Mohamed, and Hemant Misra. 2021. Multi-class text classification using bert-based active learning. *arXiv preprint arXiv:2104.14289*.

Marko Sahan, Vaclav Smidl, and Radek Marik. 2021. Active learning for text classification and fake news detection. In *2021 International Symposium on Computer Science and Intelligent Controls (ISCSIC)*, pages 87–94. IEEE.

Christopher Schröder, Andreas Niekler, and Martin Potthast. 2021. Uncertainty-based query strategies for active learning with transformers. *CoRR*, abs/2107.05687.

Simon Tong and Daphne Koller. 2001. Support vector machine active learning with applications to text classification. *Journal of machine learning research*, 2(Nov):45–66.