HW1: A brief review on Active Learning

Archisman Ghosh

CSE 584.

Contributing authors: apg6127@psu.edu;

Abstract

In this report, we present a brief review of active learning principles and strategies. Active learning involves an oracle (human annotator) which when queried in the form of unlabeled data, provides labels to the specific queries facilitating the training procedure of the machine learning algorithm. We discuss a general overview of active learning ideas and delve deep into three research papers discussing a strategy, an example, and an analysis of active learning.

Keywords: Active Learning, Information Extraction, Machine Learning.

1 Introduction

The main idea of active learning is to allow the machine learning model to choose the data from which it learns, to perform better. This is useful in cases where the data for training a model is abundant but is very difficult, and often expensive to label. In cases of Speech Recognition labeling requires the need of trained linguistic professionals and for classification-based tasks, labeling relevant data is extremely tedious. Hence, active learning systems query human annotators (oracles) to label specific unlabeled data, thus lowering the cost of obtaining labeled data [1]. There are several querying strategies like picking data points where the model is least confident in predicting or even implementing multiple models to pick data where they disagree the most. In this report, we discuss three main ideas in active learning.

1.1 Problem definition

The first problem we discuss employs active learning strategies in a situation where the data instances are grouped into bins and the bins are labeled but the individual data is not [2]. The next problem focuses on active learning in NLP (Natural Language

Parsing) [3]. The last problem discusses the annotation burden in sequence labeling and attempts to reduce it [4].

2 Multi-Instance Active Learning

Multi-Instance (MI) learning involves grouping the instances into "bags", which contain multiple instances. MI representation can be used in several machine learning tasks including content-based image retrieval and text classification.

2.1 Motivation

A characteristic of "bags" in MI learning is that a bag is labeled negative if all instances are negative, but labeled positive if any one instance is labeled positive. This raises an ambiguity thus underscoring the need to label every instance, which is often expensive and time-consuming. Hence, the paper aims to introduce active learning strategies to label particular instances to reduce the overall cost of having to label all instances and improve the training of the MI learning model.

2.2 Proposed Solution

The authors propose a framework to selectively query unlabeled instances from positive bags, to reduce ambiguity raised due to negative and positive instances being present in the same bag. In probabilistic classifiers, the uncertainty of a class label is sampled to selectively query the ones with maximum uncertainty. For an MI setting, the authors define Multi-Instance Uncertainty (MIU) as the derivative of the bag output w.r.t to the instance output times the uncertainty of the instance:

$$MIU(B_{ij}) = \frac{\partial o_i}{\partial o_{ij}} U(B_{ij})$$

where, $o_{ij} = P(y_{ij} = 1|B_{ij})$ measuring the probability that the instance B_{ij} is positive, and $U(B_{ij})$ is the uncertainty of the instance provided by the Gini measure $U(B_{ij} = 2o_{ij}(1 - o_{ij})$.

They also discuss another query strategy, Expected Gradient Length (EGL) to identify the instance with the most impact on the outcome if its label was known. EGL is calculated as

$$EGL(B_{ij}) = o_{ij}||\nabla E^{+}_{ij}(\theta)|| + (1 - o_{ij})||\nabla E^{-}_{ij}(\theta)||$$

where, $\nabla E(\theta)$ is the gradient of E w.r.t θ which is defined as a vector of partial derivatives of E with respect to the model parameters, $\nabla E(\theta) = [\frac{\partial E}{\partial \theta_1}, \frac{\partial E}{\partial \theta_2}, ..., \frac{\partial E}{\partial \theta_m}]$, and the $+$, and $-$ signs on the gradient denote the gradients for positive and negative labels respectively. The authors also state that the strategies discussed involve class probabilities to determine expected labels for instances and do not consider MI bias. They try to minimize the error over unbalanced instances rather than maximizing the expected learning model change. These strategies are integrated into multiple-instance logistic regression (MILR) models to predict the instance-level probabilities

and combine them using a softmax function and an improvement in CBIR and text classification tasks is demonstrated.

2.3 Contributions

Major contributions of the idea include

- Introducing two active learning strategies, MIU and EGL (discussed in Section 2.2) to selectively query unlabeled instances and aid in training the MI learning model.
- Testing the efficacy of their idea against the state-of-the-art baselines: using the model uncertainty and picking random instances, on CIBR tasks on the SIVAL dataset [5] and text classification tasks on the 20 Newsgroups dataset [6].
- This is the *first* work to discuss active learning strategies for MI learning setups and explores the idea of learning from labels at mixed levels of granularity (both "bag" and instance level).

2.4 Drawbacks

The ideas of EGL and MIU are mathematically complex and involve a significant computational overhead which raises scalability concerns for the strategies for larger datasets.

- The computation of $\nabla E(\theta)$ in EGL involves both $\nabla E^+(\theta)$ and $\nabla E^-(\theta)$. This can become expensive for large datasets as the computation needs to be done for each query instance. Gradient calculation happens in $O(m)$ time for m parameters, which increases significantly upon increasing the size of the dataset (the value of m becomes large).
- In MIU, there is a potential for query overlap in the same bag as it focuses on querying from the positive-labeled bag. Querying similar instances from the same bag increases the overhead of computation without contributing much to the learning of the MI model. Also, querying instances with high o_{ij} value may not yield much as the softmax function computes the derivative $\frac{\partial o_i}{\partial o_{ij}}$ which tends to zero for larger values of o_{ij}.

3 Active Learning for NLP and IE

This paper addresses the challenge of reducing annotation costs in natural language tasks. It explores active learning methods to minimize the number of labeled examples required to achieve high accuracy, focusing on two non-classification tasks: semantic parsing and information extraction (IE).

3.1 Motivation

Supervised learning in natural language processing often requires large amounts of annotated data, which is expensive and time-consuming to generate. On the other hand, unlabeled data is abundant. Active learning strategies selectively choose the most informative examples for annotation, and offers a solution to this problem by reducing the need for extensive labeled datasets. The main goal is to reduce the time

and effort required for annotation in complex tasks such as semantic parsing (mapping sentences to logical representations) and IE (extracting structured information from text). Existing applications of active learning had been limited to simple classification tasks like part-of-speech tagging, but this paper explores its utility in more complex tasks where the annotation output is intricate and costly. The reduction of overhead in annotation makes active learning a good approach for NLP tasks.

3.2 Proposed Solution

The paper introduces CHILL and RAPIER for certainty-based active learning. CHILL uses Inductive Logic Programming to construct a Shift-Reduce Parser and sets its rules from annotated examples. Hence, CHILL is used for the semantic parsing as it takes parsing as a control problem where it learns rules that determine the correct actions. Certainty metrics are calculated based on the coverage of the rules introduced. If a sentence is unparsable, the sentence with the lowest success rate is selected. RAPIER on the other hand is a rule-learning system that identifies relevant pieces of information and fills pre-defined data templates. This is hence used for IE tasks. The learning is done in a bottom-up relational manner focusing on the regular expressions of the text, including constraints like keywords and Part-of-Speech (POS) tags and semantic classes for a finer-grained learning. CHILL was trained on a subset of 250 geographical examples with Prolog queries and RAPIER on 300 job postings, and later evaluated on the F-score that combines Recall and Precision values of the prediction results.

3.3 Major Contributions

Major Contributions of the idea include

- One of the first papers to implement active learning strategies in NLP and IE tasks.
- Introduces RAPIER-a bottom-up relational parser to handle IE tasks efficiently using regular expressions of word tokens. Also implements CHILL for semantic parsing and performing non-classfication tasks on NLP data.
- Experimental data on the datasets to train CHILL and RAPIER provide a reduction of overhead by $\sim 44\%$ in annotation tasks.

3.4 Drawbacks

Although innovative, the idea of using CHILL and RAPIER have significant drawbacks that need to be addressed.

- In active learning, committee-based approaches have theoretical advantages, but they are not looked into in detail. The paper focuses on certainty-based methods only. Committee-based methods should be explored in semantic parsing and IE.
- The certainty metrics used are ad-hoc and provide a coarse-grained representation of the data complexity. Probabilistic metrics may be explored to obtain better estimates and improved sample selection.
- CHILL and RAPIER both learn on fixed batch sizes which may be detrimental to the learning of the model. An adaptive approach of setting batch size could be better if based on the current learning state of the model.

4 Active Learning for Sequence Labeling

This paper explores the effectiveness of different active learning strategies for sequence labeling tasks in NLP, presenting a comprehensive overview of existing query strategies on IE and document segmentation tasks.

4.1 Motivation

Sequence labeling tasks like IE, part-of-speech tagging, and document segmentation often require large amounts of labeled data. Acquiring this labeled data is expensive and time-consuming, making it a prime candidate for active learning, where the model selects the most informative examples for labeling. While active learning for classification tasks has been extensively studied, there is comparatively less work on active learning for sequence labeling tasks. The goal of this paper is to investigate and propose strategies that reduce the labeling effort in these tasks without sacrificing performance. The motivation arises from the need to reduce annotation costs, especially in domains like biomedical text processing, and speech recognition, where labeling is costly but unlabeled data is abundant.

4.2 Proposed Solution

The paper proposes Conditional Random Fields (CRFs) which is a probabilistic graphical model used for labeling sequential data. CRFs are more flexible than Markov Models as they eliminate the bias problem. For active learning ideas, they explore uncertainty sampling with Least Confidence (LC):

$$\phi_{LC}(x) = 1 - P(y^*|x;\theta)$$

where, ϕ_{LC} is the informativeness score for input x udner LC strategy, P being the posterior probability of the most likely label sequence y^* given x as input and θ as parameters of the model. Margin (M) is also explored as a strategy:

$$\phi_M(x) = -(P(y_1^*|x;\theta) - P(y_2^*|x;\theta))$$

, where ϕ_M is the informativeness score for x under Margin strategy and $P(y_1^*|x;\theta)$ being posterior probability for most likely sequence y_1^* and $P(y_2^*|x;\theta)$ being posterior probability for the second most likely sequence y_2^*. Token Entropy (TE) is also explored as a strategy for uncertainty sampling:

$$\phi_{TE}(x) = -\frac{1}{T}\sum_{t=1}^{T}\sum_{m=1}^{M} P_\theta(y_t = m) log P_\theta(y_t = m)$$

where, $\phi_{TE}(x)$ is the informativeness score for TE, T is the length of the sequence, and $P_\theta(y_t = m)$ is the marginal probability of label at position t in sequence x being equal to m, according to model parameter θ; and $log P_\theta(y_t = m)$ being the logarithm of marginal probability. The authors also explore Query by Committee (QBC) methods

such Vote Entropy and other strategies like Expected Gradient Length (EGL) and Information Density (ID).

4.3 Major Contributions

Major contributions of the paper include

- The paper introduces new query strategies like Sequence Entropy (SE):

$$\phi_{SE}(x) = -\sum_{y} P(y|x;\theta) log P(y|x;\theta)$$

  and Information Density, that combines SE to measure how much of the representation of x is unlabeled:

$$\phi_{ID}(x) = \phi_{SE}(x) \times \left(\frac{1}{U} \sum_{u=1}^{U} sim(x, x^{(u)}) \right)^{\beta}$$

  This takes into account the traditional uncertainty parameters along with sequence-level entropy and representativeness.
- The paper provides a thorough empirical comparison of 15 active learning strategies using eight benchmark datasets including well-known datasets like CoNLL-03 and NLPBA for named entity recognition, and BioCreative for biomedical text. Performance is measured using F1 score and area under the learning curve, with up to 150 query rounds.
- The authors show that ID and Sequence Vote Entropy (SVE) :

$$\phi_{SVE}(x) = -\sum_{y \in N_C} P(y|x;C) log P(y|x;C)$$

  perform better than LC and M. ID is particularly useful when applied to large datasets, as it considers both informativeness and representativeness and its mathematical model allows the selection of diverse and high-value sequences for labeling, reducing annotation cost.

4.4 Drawbacks

The paper proposes the idea of SVE being better. However,

- There is an overemphasis on sequence length in strategies like SVE and SE. TE clearly performs better on shorter sequences due to the normalization factor. Choosing SVE or SE for shorter sequences can lead to sub-optimal performances. Again setting up adaptive strategies to adjust to sequence length biases could be a viable solution.
- Hybrid query strategies were not explored in this paper. A combination of QBC and uncertainty may provide better solutions but was not touched upon in this work.

References

[1] Settles, B.: Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison (2009)

[2] Settles, B., Craven, M., Ray, S.: Multiple-instance active learning. Advances in neural information processing systems **20** (2007)

[3] Thompson, C.A., Califf, M.E., Mooney, R.J.: Active learning for natural language parsing and information extraction. In: Proceedings of the Sixteenth International Conference on Machine Learning. ICML '99, pp. 406–414. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1999)

[4] Settles, B., Craven, M.: An analysis of active learning strategies for sequence labeling tasks. In: Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, pp. 1070–1079 (2008)

[5] Rahmani, R., Goldman, S.: SIVAL Dataset. http://www.cs.wustl.edu/accio/. Accessed: 2024-09-06

[6] Rennie, J.: 20 Newsgroups Dataset. http://people.csail.mit.edu/jrennie/20Newsgroups/. Accessed: 2024-09-06