Jenna Warren
Connor Flynn
Dor Rubin

Project Summary:
Effects of query strategy and number of labeled samples on classification rates in active learning

The influx of data has made machine learning applications ubiquitous. While data is often readily available, in certain domains, pinpointing quality data and labeling said data can be costly. This cost has given rise to what is called active learning. Active learning is a form of supervised learning in which the training of a machine learning classifier is done on a dataset where the algorithm selects specific data points to be labeled. Despite restricting the number of labeled data points used in training, algorithms using active learning techniques can maintain or surpass accuracy in terms of performance metrics such as correct classification rate (CCR).[1]

In 2009, as part of the Pascal2 challenge program, Clopinet Labs hosted an active learning challenge in which contestants were asked to implement an algorithm using the classifier of their choice to make predictions on sample data sets. All the training data given to the contestants was unlabeled except for one point. The participants could "purchase" additional labels for a fixed cost. This cost incentivized the participants to wisely select which labels they wanted to request. Despite the ability to also use active learning in a semi-supervised approach, the majority of winning teams were able to achieve higher performance through a strictly supervised learning approach.[2]

Our research will investigate how the number of labeled data points affects metrics such as the receiver operating characteristic curve (ROC) and CCR. We will use a passive learning implementation which has access to all training labels from the start as a baseline comparison. When examining the effect of the number of labeled samples on the performance of the classifier, we will look at both the number of labeled samples in relation to the test performance, as well how different querying strategies affect performance. We hypothesize that more advanced querying techniques improve performance in fewer labelled samples than a random querying technique. Likewise, we expect to see a convergence or even surpassing of correct classification performance as the number of samples increases with that of training on the entire set.

In reference to the competition, we will use a set of classifiers and datasets that align with the top performing strategies. Using the IBN_SINA dataset, we take on the task of spotting arabic words in an ancient manuscript.[3] The data is of a mixed feature type thus requiring modest preprocessing. It is represented by 92 variables with approximately 81% sparsity. The initial training set is comprised of roughly 10,000 data points while the final test set will be in the range of

[1] Settles, Burr. "Active Learning Literature Survey." Computer Sciences Technical Report 1648.
[2] I. Guyon et al. Datasets of the active learning challenge. Technical Report, 2010.
[3] IBN_SINA IBN SINA: http://www.causality.inf.ethz.ch/al_data/IBN_SINA.html

25,000. The competition provided a preliminary toy dataset called ALEX consisting of 5,000 training and a separate 5,000 test samples to prototype the development.

In correspondence to the strategies used by the top teams in the competition, we will employ a linear kernel ridge regression, naive bayes and decision trees with a passive learning implementation. Our fallback will be to look exclusively at linear kernel ridge regression and decision trees since they were the most common strategies in the competition. The performance of each classifier will be analyzed and set as the benchmark for the remainder of the study. Next, we will use the querying strategies detailed below to analyze their effect in combination with the number of labeled data points on each of the above base classifiers.

There are three main active learning scenarios: membership query synthesis, stream-based selective sampling and pool-based sampling. Due to its effectiveness we will limit our research to the pool-based approach where the learner can choose from the entire pool for labeling. Given additional time, we will investigate the stream-based approach where for each arrived point, the learner decides whether to query for the label or discard the data point.

The ideal goal will involve testing the classifiers using three different query strategies: random querying, uncertainty sampling, and query by committee. [4]Notably, all querying will involve picking data already listed in the training/testing sets and not de novo querying, which involves inventing new feature vectors to query. Unless necessary for computation speed, batch querying will be avoided in favor of querying one point at a time.

In this context, to query a point means to ask for the oracle for the label belonging to a feature vector. As the name suggests, random querying randomly picks data from the set to be labeled by the oracle; no metric needs to be used to decide which points will be queried. Uncertainty sampling uses some measure of uncertainty and queries the points in which the model is most uncertain. The metric to be used for uncertainty sampling in this project will be information entropy[5]. In other words, the point to be queried will be the one in which the information entropy is at a maximum.

Query by committee presents the most challenge and will only be present in the ideal objective. Query by committee involves having more than one competing model all evaluate the unknown feature vectors in the set and the point whose label they disagree on the most is queried. This may be achieved by averaging the KL divergence[6] of the label distributions of the individual committee members compared to the aggregate committee and taking the point which maximizes this (for generative models). Another querying metric for this strategy is a measure called the vote entropy[7] which is like information entropy only instead of using the posterior of a label given x, we use the average number of votes for a point across the committee. More study needs to be put into

[4] Settles, Burr. "Active Learning Literature Survey." Computer Sciences Technical Report 1648.

[5] Formula - A

[6] Formula - B

[7] Formula - C

these metrics in order to evaluate which one will be used, however the most likely candidate is vote entropy both due to it being objectively simpler than the average KL divergence and it being closely related to the proposed metric for uncertainty sampling (information entropy).[8]

Finally, in the competition summary report, the authors describe that “Interestingly no participant used Bayesian active learning”. One of our fallback options is to explore why no team elected to go that route and explore whether it is a potential fit for our own query strategy.

Division of Labor:

The primary point person for each area of this project will be as follows:

- Coding: Dor Rubin
- Presentation / Writing: Jenna Warren
- Research: Connor Flynn

Formulae:

A. Information Entropy:
   - $x_{query} = argmax\ [- \sum_{i} P_{\theta}(y_i || x) log(P_{\theta}(y_i || x))$

B. KL Divergence for Committee Querying:
   - $x_{query} = argmax\ \frac{1}{C} \sum_{c=1}^{C} D(P_{\theta(c)} || P_C)$
   - Where $P_{\theta(c)}$ is the label distribution of a given committee member, P_C is the aggregate label distribution of the full committee, C is the committee size.

C. Voting Entropy:
   - $x_{query} = argmax\ [- \sum_{i} \frac{V(y_i)}{C} log(\frac{V(y_i)}{C})$

[8] Settles, Burr. "Active Learning Literature Survey." Computer Sciences Technical Report 1648.

Works Cited:

Cohn, Z. Ghahramani, and M.I. Jordan. Active learning with statistical models. Journal of Artificial Intelligence Research, 4:129–145, 1996.

G. C. Cawley. Some baseline methods for the active learning challenge. Journal of Machine Learning Research, Workshop and Conference Proceedings (in preparation), 10, 2010. D.

G. Paass and J. Kindermann. Bayesian query construction for neural network models. In Advances in Neural Information Processing Systems (NIPS), volume 7, pages 443–450. MIT Press, 1995.

I. Guyon et al. Datasets of the active learning challenge. Technical Report, 2010. URL http://clopinet.com/al/Datasets.pdf. Vincent Lemaire and Christophe Salperwyck. Post-hoc experiments for the active learning challenge. Technical report, Orange Labs, 2010.

IBN SINA

http://www.causality.inf.ethz.ch/al_data/IBN_SINA.html

Settles, Burr. "Active Learning Literature Survey." Computer Sciences Technical Report 1648. University of Wisconsin-Madison, 26 Jan. 2010. Web. 25 Mar. 2017. <http://burrsettles.com/pub/settles.activelearning.pdf>.