ACTIVE LEARNING PROJECT 1 TECHNICAL REPORT

Exploring Active Learning Strategies and Classifiers on Sklearn Datasets

Implemented By:

Arwa Sallam	20200067
Natalie Monged	20200586
Mahmoud Wael	20200505
Aly Walid	20200336

MARCH 25, 2024

Table of Contents

List of Figures

Glossary

- US=Uncertainty Sampling.
- RS=Random Sampling
- QBC=Query by Committee.
- LP=Label Propagation Classifier.
- SVC=Support Vector Classifier.
- PW=Parzen Window Classifier.
- MM=Mixture Model Classifier.
- E=Entropy method.
- MS=Margin Sampling method.
- LC=Least Confident method.
- KL Divergence=Kullback-Leibler divergence method.
- VE=Vote Entropy method.
- USxE=After/By using Entropy method with Uncertainty Sampling strategy.
- USxMS=After/By using Marginal Sampling method with uncertainty sampling strategy.
- USxLC=After/By using Least Confident method with uncertainty sampling strategy.
- QBCxKL=After/By using KL Divergence method with Query by Committee strategy.
- QBCxKL=After/By using Vote Entropy method with Query by Committee strategy.

Abstract:

Active machine learning is a powerful approach that iteratively selects the most informative data points for labeling, thereby reducing the labeling effort while maintaining or improving the model's performance. Active machine learning addresses this challenge by actively selecting the most informative instances for labeling, thereby maximizing the performance gain with minimal labeling effort. In this report, we present our implementation of active machine learning on various datasets using a combination of classifiers and uncertainty sampling strategies.

Introduction:

Conventional supervised learning requires large data sets with supervised labels to form perfect models. This involves labelling many data sets which can be costly and time-consuming. In this technical report, we implement active machine learning in different scikit-learn datasets such as Iris, Breast Cancer, Wine, and Diabetes using multiple classifiers and uncertainty along with query by committee sampling techniques. Additionally, we explore the challenges posed by imbalanced datasets, explicitly by our utilization of the Health Care Stroke dataset. The classifiers in our implementation include a Label Propagation, Support Vector Classifier (SVC), Parzen Window and a Mixture Model that works to solve this problem by coming up with extra labels. The uncertainty sampling strategies include entropy, margin sampling and least confident and query by committee strategies such as vote entropy and kl_divergence. We aim to show through various approaches with active machine learning we can generate form a model with a higher f1score than that of a conventional classifier while lessening the amount of labeled data needed.

Implementation Design:

- **Datasets Selection:** We utilized four widely used datasets for classification and regression tasks: Iris, Breast Cancer, Wine, Diabetes, and the imbalanced Health Care Stroke.

- **Classifier Selection:** We employed multiple classifiers to evaluate the performance of active learning. These classifiers include Label Propagation(LP), Support Vector Classifier (SVC), Parzen Window(PW), and Mixture Model(MM).

- **Uncertainty Sampling (US) Strategies:** We applied uncertainty sampling strategies to select the most informative instances for labeling. The strategies include Entropy (E), Margin Sampling (MS), and Least Confident (LC).

- **Query by Committee Strategies (QBC):** In addition to uncertainty sampling, we incorporated query by committee strategies such as Vote Entropy (VE) and KL divergence to further enhance the selection of informative instances.

- **Active Learning Approach:** As a result of our confusion, we worked on 2 approaches; the first one is to consider all the data unlabeled (NaN) following scikit-active ml documentation notebook and the other one is to consider the data with a majority of unlabeled (NaN) targets following the lecture's and the lab's explanation.

- **Active Learning Cycle:** We implemented an active learning loop that iteratively selects instances for labeling based on the chosen sampling strategies. After each iteration, the selected instances are labeled, and the classifier is retrained on the updated labeled dataset.

- **Evaluation Criteria:** We applied metrics like F1 Score and Confusion Matrix which were sufficient for comparison as shown below in the next section. Moreover, we applied F1 Score and Confusion Matrix with any US method and just F1 Score with any QBC method.

1st Approach Comparison:

- N.B: "We ran each combination Twice for validity"
- For IRIS dataset (N.B Values represent f1 score ratio with # Cycles=25)

  - Label Propagation Testing F1 Score (Before AL Cycle)

    - QBCxKL: 0.367
    - QBCxVE: 0.365
    - USxE: 0.46
    - USxMS: 0.428
    - USxLS: 0.378

  - Label Propagation Testing F1 Score (After AL Cycle)

    - QBCxKL: 0.81
    - QBCxVE: 0.794

- ❖ USxE: 0.923
- ❖ USxMS: 0.92
- ❖ USxLS: 0.9

- For Breast Cancer Dataset (N.B: Values represent f1 score ratio with # Cycles=25)
  - SVC Testing F1 Score (Before AL Cycle)
    - ❖ QBCxKL: 0.595
    - ❖ QBCxVE: 0.545
    - ❖ USxE: 0.544
    - ❖ USxMS: 0.57
    - ❖ USxLS: 0.56

  - SVC Testing F1 Score (After AL Cycle)
    - ❖ QBCxKL: 0.722
    - ❖ QBCxVE: 0.84
    - ❖ USxE: 0.892
    - ❖ USxMS: 0.9
    - ❖ USxLS: 0.9

1st Approach Conclusion:

After inspecting our background runs and comparison section, we landed our decision on IRIS and Breast Cancer datasets as they had the best performance due to the massive jump in the classifiers' f1 score after applying the active learning cycle so that makes them the 1st two members of our active learning model.

Moving on to the classifiers and strategies in IRIS, we decided to favor Label Propagation due to its robustness and stability across multiple strategies and for the strategies we decided to use Entropy method as it gave the highest f1 score.

On the other hand, it wasn't that confusing to choose the best pair of classifier and strategy in the Breast Cancer dataset. Obviously, our winning pairs were SVC classifier with Marginal Sampling method in the Uncertainty Sampling Strategy and SVC classifier with Marginal Sampling method in the Uncertainty Sampling Strategy Least Confident, but we decided to favor the 1st pair.

Finally, for more clarification you can see below 2 figures visualizing IRIS and Breast Cancer datasets along with a pool of strategies with Label Propagation and SVC classifiers respectively.

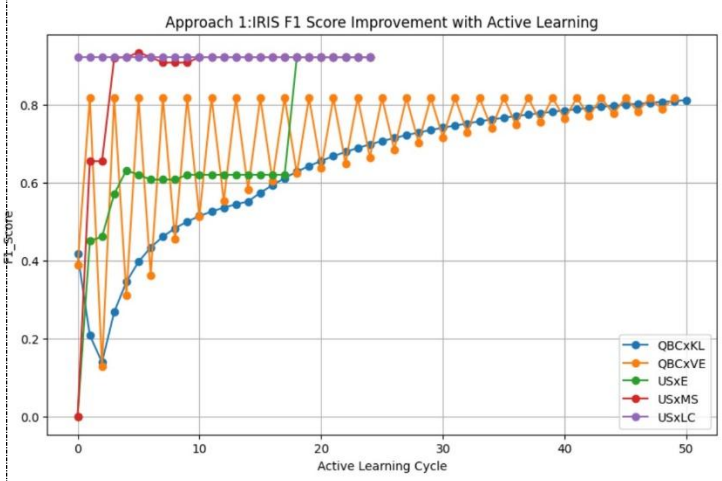

Figure 2-Appr1 IRIS F1 Score Improvement

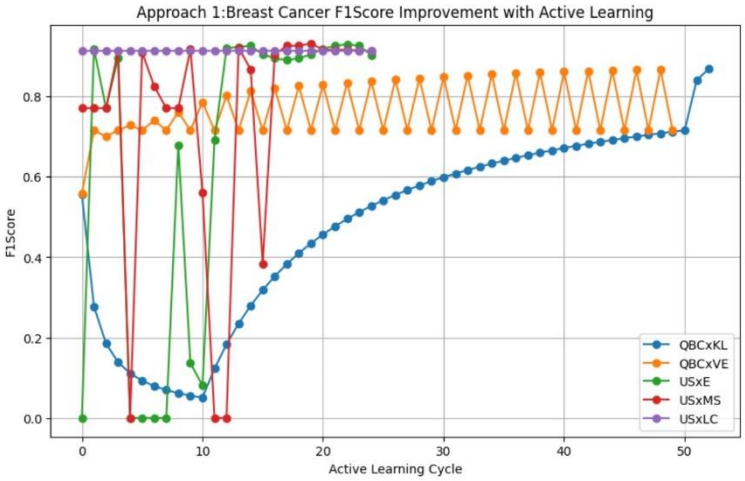

Figure 2-Appr2 Breast Cancer F1 Score Improvement

2nd Approach Comparison:

- For Wine dataset (N.B: Values represent f1 score with # Cycles=25, Batch Size=10)
  - Parzen Window f1 score (Before AL Cycle)
    - QBCxKL: 0.1
    - RS: 0.22
    - USxMS: 0.23
    - USxE: 0.28

- Parzen Window f1 score (After AL Cycle)
    - QBCxKL: 0.71
    - RS: 0.7
    - USxMS: 0.68
    - USxE: 0.72

- For Breast Cancer Dataset (N.B: Values represent f1 score ratio with # Cycles=25, Batch Size=13)
    - SVC Accuracy (Before AL Cycle)
        - QBCxKL: 0.73
        - RS: 0.7
        - USxMS: 0.74
        - USxE: 0.73
    - SVC Accuracy (After AL Cycle)
        - QBCxKL: 0.92
        - RS: 0.93
        - USxMS: 0.928
        - USxE: 0.923

2nd Approach Conclusion:

Due to multiple combinations between the parameters and techniques and thus test cases complexity we provided a summary of our running experiences with the different datasets along with strategies and classifiers in the Comparison Section.

As a result of inspecting them, eventually we landed on Wine and Breast cancer as they yielded the best performance among the datasets.

Accordingly, we chose SVC and RS to be our winning pair with Breast Cancer dataset even though it went through very tough competing results with the rest of the strategies as we couldn't neglect their performance as well that can be used without affecting the overall performance that much.

Finally, we chose Parzen Window and Entropy to be our winning pair with Wine dataset as they provided the most stable and robust performance. You can look at the figures below to observe our choices.

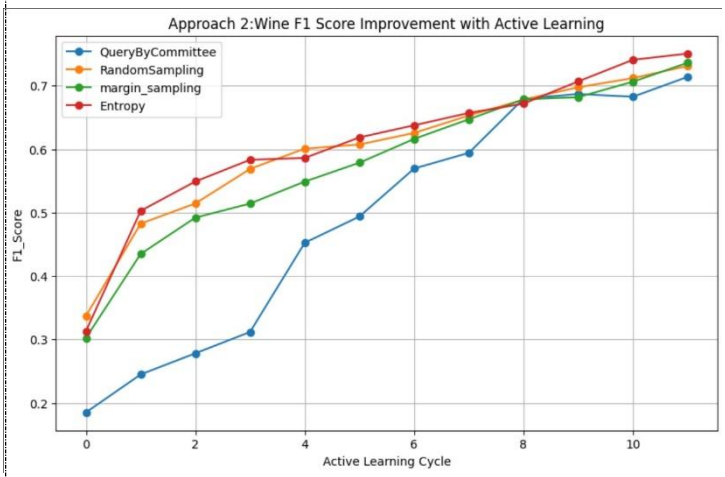

Figure 4-Appr 2 Wine F1 Score Improvement

Figure 4-Appr2 Breast Cancer F1 Score Improvement

Imbalanced Dataset Comparison:

- For Healthy Stroke dataset (N.B: Values represent f1 score with # Cycles=16, Batch Size=230)
  - Parzen Window f1 score (Before AL Cycle)
    - ❖ USxE: 0.01
    - ❖ USxLC: 0.02
    - ❖ RS: 0.02
    - ❖ QBCxKL: 0.03

- Parzen Window Ac Parzen Window f1 score curacy (After AL Cycle)
  - ❖ USxE: 0.04
  - ❖ USxLC: 0.045
  - ❖ RS: 0.044
  - ❖ QBCxKL: 0.05

Imbalanced Dataset Conclusion:

Due to multiple combinations between the parameters and strategies and thus test cases complexity we provided a summary of our running experiences with the different strategies along with classifiers in the Comparison Section.

As a result of inspecting them, eventually we landed on Parzen Window to be our classifier as it yielded the best performance among the classifiers.

Within the framework of that, we witnessed another tough race between the strategies as provided in the comparison section, but we decided to choose the KL divergence due to the relatively high jump of f1 score difference before and after the active learning cycle.

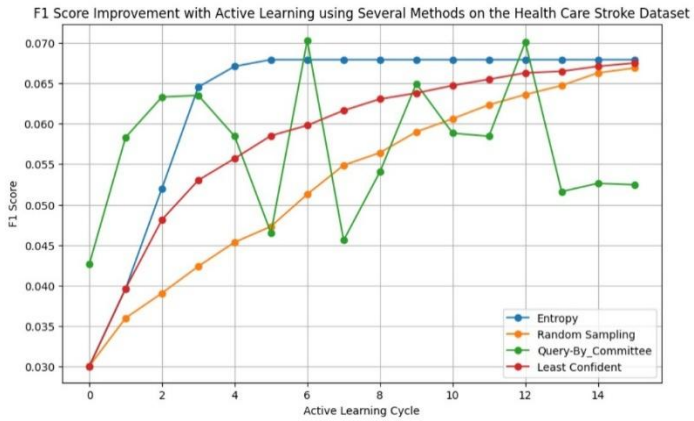

Figure 6-HCS F1 Score Improvement

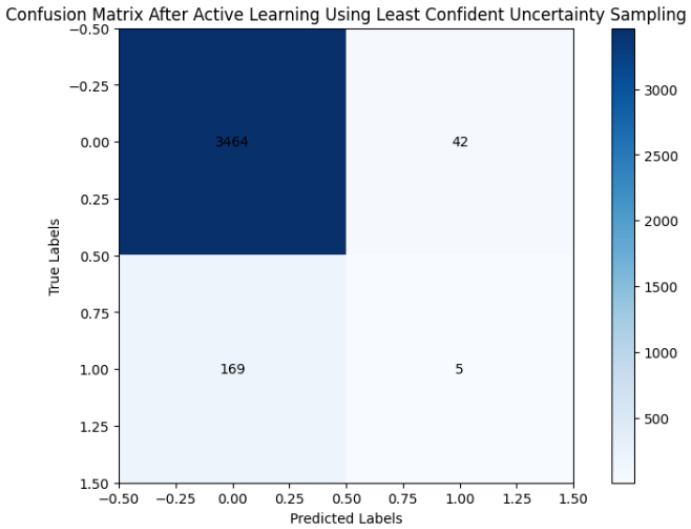

Figure 6-HCS CF After AL

References:

ML | Active Learning - GeeksforGeeks

scikit-activeml · PyPI

scikit-activeml/tutorials/00_pool_getting_started.ipynb at master · scikit-activeml/scikit-activeml (github.com)