Active Machine Learning - Miniproject

02463 Active machine learning and agency

Lucas Emcken
Technical University of Denmark
s214625@student.dtu.dk

Lucas Sylvester
Technical University of Denmark
s214636@student.dtu.dk

William Peytz
Technical University of Denmark
s204145@student.dtu.dk

[ ]

1 ABSTRACT

Acquiring accurately labeled data for training machine learning models can be time-consuming and expensive, posing challenges in various domains, including fraudulent credit card transactions. This project explores the potential of active machine learning, specifically query by committee (QBC) and uncertainty sampling, in improving the convergence speed and accuracy compared to random sampling. The study focuses on a pool-based active learning scenario using a logistic regression model to categorize credit card transactions as valid or fraudulent. Five permutations of the dataset are evaluated using different query strategies, starting with random initial points and iteratively selecting points based on the chosen strategy. The accuracy of the trained models is tested on separate test datasets. The results demonstrate that both least confidence uncertainty sampling and query by committee outperform random sampling, with least confidence achieving the highest mean accuracy. Furthermore, these active learning methods exhibit lower standard deviation, indicating more consistent model performance across multiple runs. It is acknowledged that different query strategies may yield better results depending on the specific scenario and dataset. Additionally, it should be noted that QBC requires additional training time compared to other methods.

2 INTRODUCTION

Training machine learning models often requires extensive and accurately labelled data, which is not always readily available. It might be either extremely time consuming or very expensive to acquire. A prominent example of this is fraudulent credit card transactions. Even thought there are enormous amounts of data available, its is very expensive and time consuming to label, as we will need to hire people to individually call and confirm transactions. We will therefore investigate if there is a case for using active machine learning for selecting what transactions are worth investigating to train on. In this project we will use query by council and uncertainty sampling to see if it leads to better and faster convergence of accuracy compared to random sampling. In this project we will consider this as a database of transactions, so the pool-based scenario of active learning.

3 METHODS

The direct problem that we want our models to optimize is categorising card transactions as either valid or fraudulent. To do this we will be using a logistic regression model and the described transaction data. To evaluate the different query strategies and the baseline, 5 permutations of the dataset was made. For each of the permutations an initial random 20 points was selected. The different query strategies then iteratively selected a point, the corresponding classifier then trained on that point and then the accuracy was tested on the permutations test dataset. This was repeated until a 100 points was selected. This was repeated for each of the 5 permutations. The mean score and standard deviation for each number of points and strategy was then calculated to see if there was a significant difference between the strategies.

Data

For this project, we will be working with credit card fraud data[1], the dataset contains 1 million points, of which 87,403 points are fraudulent. We will be using a reduced dataset by sampling 10,000 points in the interest of code execution time. It should be noted that this data likely is synthetic, as there is no units attributed to the values, and users were able to achieve a very high accuracy using simple tree classifiers. Regardless, this project is not about detecting credit card fraud, but rather is about machine learning methods, so it should not affect the results or learning outcome in any major way. The models trained have also have their predictive power balanced, such that a baseline guess of only picking one class would return 50% accuracy. The dataset contains data with an exponential distribution and as such has been log transformed, and scaled such that it now follows the normal distribution $X \sim \mathcal{N}(0, 1)$

Logistic Regression

For this project, we will use logistic regression as our model. Logistic regression is a type of statistical analysis used to model the relationship between a categorical dependent variable and one or more independent variables. It does this by using a logistic function, which maps an input to a value between 0 and 1. The output of the logistic function can be interpreted as the probability of the dependent variable taking on a particular value, given the values of the independent variables.

Query strategies

Query strategies is a useful machine learning tool for active learning. Sampling through query strategies allows a machine learning algorithm to pick out data points, which might provide the most information, and use them for training. This is useful when labeling data is expensive (either time or money). For this project, we will be uncertainty sampling with Least confidence, vote entropy via query by committee, and comparing them to a baseline random sampling.

Baseline - Random sampling A baseline query strategy that picks a new random sample from the pool of unlabeled data.

Uncertainty Sampling Uncertainty sampling uses the current models uncertainty, and therefore requires a model that provides such. Since we have binary classification problem with a probabilistic model, the uncertainty is the difference between the most confident prediction and 100% confidence. This query strategy samples the point with the least confidence. Other variations include margin sampling (difference between two most likely points) and entropy (distribution).

Query-By-Committee and Vote Entropy Query-By-Committee (QBC) uses multiple models trained on the dataset, and their predictions are combined to form the final output. We will be using the committee to calculate the entropy of each point, and sample the point with the largest entropy between all predictions, i.e. the point which the committee disagrees on the most, for training. Our committee consists of 5 members.

4 RESULTS

Training the models over 15 repeats of querying 100 points yields the following accuracies.

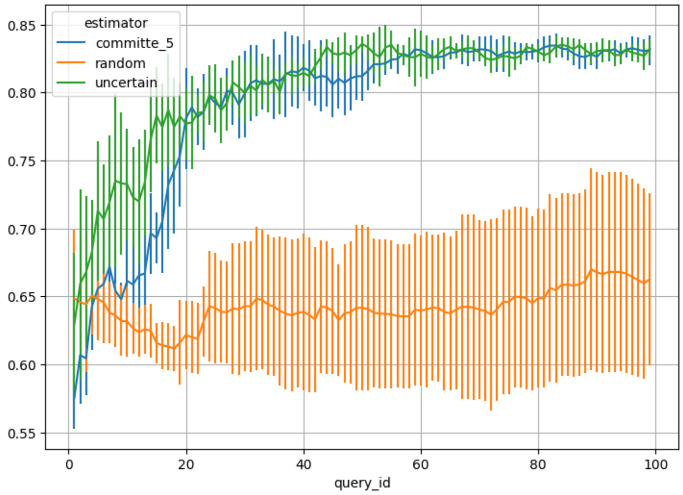

Figure 1: Balanced accuracy of Random sampling, Query by Committee and Least Confident

	mean	std
Committee	0.831	0.011
Least Confidence	0.832	0.007
Random	0.663	0.064

Table 1: Final accuracies of the models

From figure 1 and table 1 we see that Least confidence performs the best, followed by QBC and finally random sampling

5 DISCUSSION

From the plot on figure 1, there is a significant improvement by using either least confidence uncertainty sampling or query by committee compared to just random sampling. Both the mean predictive power and the standard deviation is significantly better than random sampling, showing that by using active learners, you are able to improve model performance, and achieve a more consistent model performance over multiple runs.

It is possible that other query strategies may perform better in different scenarios or different data sets. A typical problem with least confidence is the tendency to pick outliers that wont necessarily provide much relevant information. The good performance of the least confidence strategy may be because there are very few outliers in the data set, since it likely is artificially generated. Worth noting is that QBC also took significantly more time to train.

REFERENCES

[1] Dhanush Narayanan R. 2022. Credit Card Fraud Detection. https://www.kaggle.com/datasets/dhanushnarayananr/credit-card-fraud