TECHNICAL REPORT

NAME	Moustafa Omar Mohammed
ID NUMBER	20200542
NAME	Ezz El-Din Ahmed Saber
ID NUMBER	20200325
NAME	Ahmed Hany Ibrahim
ID NUMBER	20200054
NAME	Abdelrahman Ramadan Aboulela
ID NUMBER	20200284

Table of Contents

Introduction

Active machine learning techniques have gained prominence in recent years due to their ability to optimize the learning process by intelligently selecting informative instances for labeling. In this report, we delve into the utilization of active learning methodologies on three distinct datasets: dry bean, date fruit, and wine quality.

The dry bean dataset provides a rich source of information for classifying different varieties of dry beans based on various features such as shape, texture, and color attributes. Meanwhile, the date fruit dataset encompasses diverse characteristics of date fruits, presenting a unique challenge for classification tasks. Lastly, the wine quality dataset represents an imbalanced scenario, where certain classes are underrepresented, posing significant hurdles for conventional machine learning approaches.

To address these challenges, we employ active learning strategies including least confident, margin sampling, random sampling, and entropy-based techniques. By iteratively selecting the most informative samples for annotation, these methods aim to improve classification performance while minimizing labeling costs.

Through comprehensive experimentation and evaluation, we aim to elucidate the effectiveness of active learning methodologies across these datasets. Furthermore, we seek to provide insights into the suitability of these techniques for handling imbalanced data distributions and enhancing classification accuracy in real-world scenarios.

Abstract

This project explores the application of active machine learning methodologies on diverse datasets, encompassing dry bean, date fruit, and an imbalanced wine quality dataset. Leveraging techniques such as least confident, margin sampling, random sampling, and entropy-based strategies, we aim to enhance classification performance and mitigate the challenges posed by imbalanced data distribution. Through rigorous experimentation and analysis, we assess the efficacy of these active learning techniques across varied datasets, shedding light on their suitability for real-world applications.

Dataset description

Dry Bean Dataset: The Dry Bean Dataset comprises various features extracted from images of different varieties of dry beans. These features include shape, texture, and color attributes. Each instance in the dataset represents a single dry bean sample, and the goal is to classify the beans into their respective varieties based on these features. This dataset is valuable for tasks related to agricultural research, crop classification, and food quality assessment.

Date Fruit Dataset: The Date Fruit Dataset contains information on various attributes of date fruits, including size, color, texture, and sweetness level. Each entry in the dataset corresponds to a single date fruit sample. The dataset aims to facilitate the classification and characterization of different types of date fruits based on these attributes. It is useful for agricultural studies, market analysis, and quality control in the date fruit industry.

Wine Quality Dataset: The Wine Quality Dataset consists of chemical properties of red and white wines, such as acidity levels, sugar content, pH, and alcohol percentage. Additionally, it includes sensory data representing the quality ratings of the wines by experts. The dataset comprises a mix of red and white wines, with quality ratings ranging from low to high. The primary objective is to predict the quality of wines based on their chemical composition. This dataset is valuable for wine production optimization, quality assurance, and sommelier analysis.

Techniques

Uncertainty Sampling: Uncertainty sampling is a strategy that selects instances for which the model is uncertain about its prediction. The uncertainty is often measured using a measure such as entropy or variance. For a binary classification problem, entropy-based uncertainty sampling can be calculated as:

$$H(p) = -p\log_2(p) - (1-p)\log_2(1-p)$$

where *p* is the probability of the positive class.

Support Vector Machine (SVM): Support Vector Machine is a supervised learning algorithm used for classification tasks. It works by finding the hyperplane that best separates the classes in the feature space. The decision function for SVM can be represented as:

$$f(x) = \text{sign}(\mathbf{w} \cdot \mathbf{x} + b)$$

where **w** is the weight vector, **x** is the input feature vector, and *b* is the bias term.

Least Confident Sampling: Least confident sampling selects instances where the model has the lowest confidence in its prediction. It is calculated as the difference between 1 and the maximum predicted probability:

$$LC(x) = 1 - \max(p(y|x))$$

Margin Sampling: Margin sampling selects instances where the difference in predicted probabilities between the top two classes is small, indicating uncertainty. For binary classification, it can be calculated as:

$$MS(x) = p(y_1|x) - p(y_2|x)$$

where y_1 and y_2 are the top two predicted classes.

Random Sampling: Random sampling selects instances randomly from the unlabeled pool without considering model uncertainty or decision boundaries.

Entropy-based Sampling: Entropy-based sampling measures uncertainty by calculating the entropy of class probabilities. It is computed as:

$$E(x) = -\sum_{i=1}^{C} p_i \log_2(p_i)$$

where p_i is the probability of class *i* and *C* is the number of classes.

Implementation Design

1. Introduction:

Active learning is a machine learning paradigm that involves an iterative process of selecting the most informative data points for annotation to improve model performance with minimal labeling efforts. In this technical report, we present an active learning framework tailored for multi-class classification tasks using various sampling strategies.

2. Dataset Preprocessing:

The framework begins with preprocessing the datasets. Two datasets are utilized: the Dry Bean Dataset and the Date Fruit Dataset. The preprocessing steps include:

Separating features (X) and labels (y).

Scaling the features using min-max scaling.

Encoding categorical labels using LabelEncoder.

3. Active Learning Strategies:

Four active learning strategies are employed to select the most informative data points for annotation in each iteration:

Random Sampling: Randomly selects data points for annotation.

Least Confident Sampling: Select the data points where the model is least confident in its predictions.

Margin Sampling: Identifies data points with the smallest margin between the top two predicted class probabilities.

Entropy Sampling: Selects data points based on the entropy of class probabilities.

4. Model Training and Evaluation:

A Random Forest classifier is utilized as the base model within the active learning framework. The framework iteratively selects a batch of data points based on the chosen sampling strategy, labels them, and updates the model. The process continues for a fixed number of rounds.

5. Evaluation Metrics:

For each round of active learning, both training and testing accuracies are computed and recorded for analysis.

6. Implementation Details:

The framework is implemented in Python using the scikit-learn library for model training and evaluation. The SkactiveML library is leveraged for active learning functionalities, including querying and updating the unlabeled data pool.

7. Handling Imbalanced Datasets:

An additional experiment is conducted on an imbalanced dataset (Wine Quality Dataset) to evaluate the framework's performance in handling class imbalances.

8. Results and Analysis:

The training and testing accuracies obtained for each sampling strategy are recorded and analyzed to assess their effectiveness in improving model performance over iterations.

Results

Result for balanced datasets

Highest Accuracy Margin Sampling: Utilizing highest accuracy margin sampling resulted in the selection of data points where the model exhibited the greatest uncertainty, leading to focused learning on challenging instances. This method yielded superior accuracy improvements compared to other sampling strategies.

Entropy-Based Sampling: Employing entropy-based sampling prioritized data points with high uncertainty, allowing the model to learn from diverse and informative examples. The approach led to notable enhancements in accuracy by focusing on regions of the data space where the model was uncertain.

Least Confident Sampling: Least confident sampling targeted data points where the model's prediction confidence was lowest, facilitating targeted learning in regions of uncertainty. While effective, this method achieved slightly lower accuracy gains compared to other sampling techniques due to its conservative selection approach.

Random Sampling: Random sampling involved selecting data points without considering their relevance or uncertainty, serving as a baseline for comparison. While straightforward, this method yielded the lowest accuracy improvements compared to the more strategic sampling strategies, highlighting the importance of informed data selection in enhancing model performance.

Imbalanced data

Wine quality dataset got accuracy less than balanced dataset

Models trained on imbalanced data tend to favor the majority class, leading to misclassification of the minority class.

Sampling Techniques' Impact: Even with sampling techniques, minority class instances may not be adequately represented, limiting their effectiveness.

Generalization Challenges: Imbalanced data can result in models that struggle to generalize to unseen data, particularly for the minority class.

Conclusion:

The training and testing accuracies obtained for each sampling strategy are recorded and analyzed to assess their effectiveness in improving model performance over iterations.

The active learning framework demonstrates its capability to enhance model performance by selecting the most informative data points for annotation, thus reducing the labeling efforts required while maintaining competitive accuracy levels.

References

- Learning Loss for Active Learning | Papers With Code
- [2012.04225] Active Learning: Problem Settings and Recent Developments (arxiv.org)
- How to measure uncertainty in uncertainty sampling for active learning | Machine Learning (springer.com)
- Entropy-based Sampling Approaches for Multi-Class Imbalanced Problems | IEEE Journals & Magazine | IEEE Xplore
- Active learning approach using a modified least confidence sampling strategy for named entity recognition | Progress in Artificial Intelligence (springer.com)
- The uncertainty sampling (least confidence) with bias for different pp... | Download Scientific Diagram (researchgate.net)
- Improved Margin Sampling for Active Learning | SpringerLink
- Margins (apa.org)
- Entropy-based Sampling Approaches for Multi-Class Imbalanced Problems | IEEE Journals & Magazine | IEEE Xplore
- Entropy-based hybrid sampling ensemble learning for imbalanced data - Dongdong - 2021 - International Journal of Intelligent Systems - Wiley Online Library