Introduction

Active Learning in Machine Learning
Active learning is a machine learning technique that allows the model to iteratively query the user for labels on unlabeled data points. This allows the model to focus its learning on the most informative data points, which can lead to improved performance with less training data. Active learning is particularly useful in applications where labeled data is expensive or difficult to obtain.

How Active Learning Works
Active learning works by selecting the most informative data points for the model to learn from. This is typically done using a measure of uncertainty, such as the entropy of the model's predictions. The model then queries the user for the labels of these data points. Once the user provides the labels, the model updates its predictions and repeats the process.

Benefits of Active Learning
Active learning offers a number of benefits over traditional machine learning techniques, including:

- **Reduced need for labeled data:** Active learning can reduce the amount of labeled data needed to train a model, which can save time and money.
- **Improved performance:** Active learning can lead to improved performance on a variety of tasks, such as classification, regression, and clustering.
- **Reduced risk of overfitting:** Active learning can help to reduce the risk of overfitting, which occurs when a model learns too much from the training data and is unable to generalize well to new data.

Applications of Active Learning
Active learning is used in a variety of applications, including:

- **Natural language processing:** Active learning is used to train natural language processing models, such as those used for machine translation and text classification.
- **Computer vision:** Active learning is used to train computer vision models, such as those used for object detection and image classification.
- **Speech recognition:** Active learning is used to train speech recognition models, such as those used in voice assistants and dictation software.

Conclusion
Active learning is a powerful machine learning technique that can be used to improve the performance of models with less training data. Active learning is particularly useful in applications where labeled data is expensive or difficult to obtain.

Objectives

1. **Reduce the amount of labeled data needed to train a model:**
    - Active learning can help reduce the need for large amounts of labeled data, which can be time-consuming and expensive to obtain.
    - By actively selecting the most informative data points to label, active learning can achieve similar or better performance with significantly less labeled data.
2. **Improve the performance of models on a variety of tasks:**
    - Active learning has been shown to improve the performance of machine learning models on a wide range of tasks, including classification, regression, and natural language processing.
    - By focusing on the most informative data points, active learning can help models learn more effectively and efficiently.
3. **Reduce the risk of overfitting:**
    - Overfitting occurs when a model learns too much from the training data and starts to make predictions that are too specific to the training set.
    - Active learning can help reduce overfitting by selecting data points that are representative of the underlying distribution of the data, rather than just the training set.
4. **Make machine learning models more efficient and effective:**
    - By actively selecting the most informative data points to label, active learning can help make machine learning models more efficient and effective.
    - This can lead to reduced training times, improved model performance, and lower computational costs.
5. **Enable machine learning models to learn from small datasets:**
    - Active learning can be particularly useful for learning from small datasets, where traditional machine learning methods may not have enough data to learn effectively.
    - By actively selecting the most informative data points, active learning can help models make the most of the available data.
6. **Allow machine learning models to learn from unlabeled data:**
    - Active learning can be used to learn from unlabeled data, which is often available in much larger quantities than labeled data.
    - By actively selecting the most informative unlabeled data points to label, active learning can help models learn more effectively from both labeled and unlabeled data.
7. **Improve the interpretability of machine learning models:**
    - Active learning can help improve the interpretability of machine learning models by providing insights into the decision-making process of the model.
    - By actively selecting the most informative data points, active learning can help identify the features that are most important to the model's predictions.
8. **Make machine learning models more robust to noisy and incomplete data:**
    - Active learning can help make machine learning models more robust to noisy and incomplete data by selecting data points that are representative of the underlying distribution of the data.
    - This can help models learn more effectively from real-world data, which is often noisy and incomplete.

Scenarios

Pool-Based Active Learning

Pool-based active learning is a type of active learning where the model selects the most informative data points from a pool of unlabeled data. The pool of unlabeled data is typically much larger than the labeled data set, and the model iteratively selects data points from the pool to label.

How Pool-Based Active Learning Works

1. Initialize the labeled data set and the pool of unlabeled data.
2. Train a model on the labeled data set.
3. Select the most informative data points from the pool of unlabeled data.
4. Query the user for the labels of the selected data points.
5. Add the labeled data points to the labeled data set.
6. Repeat steps 2-5 until the desired performance is achieved.

Benefits of Pool-Based Active Learning

- **Reduced need for labeled data:** Pool-based active learning can reduce the amount of labeled data needed to train a model, which can save time and money.
- **Improved performance:** Pool-based active learning can lead to improved performance on a variety of tasks, such as classification, regression, and clustering.
- **Reduced risk of overfitting:** Pool-based active learning can help to reduce the risk of overfitting, which occurs when a model learns too much from the training data and is unable to generalize well to new data.

Membership Query Syntheses

Membership query synthesis (MQS) is a technique used in active learning to generate new data points that are informative for the model. MQS is based on the idea that the model can learn from its own predictions, without the need for additional labeled data.

The basic idea of MQS is to generate a set of candidate data points, and then query the model to predict the labels of these data points. The model's predictions are then used to select the most informative data points, which are then added to the training set.

There are several different methods for generating candidate data points for MQS. One common method is to use a sampling strategy, such as random sampling or stratified sampling. Another method is to use a generative model to generate new data points.

Once the candidate data points have been generated, the model is queried to predict the labels of these data points. The model's predictions are then used to select the most

informative data points. There are several different criteria that can be used to select the most informative data points, such as the entropy of the model's predictions or the expected model gain.

MQS has been shown to be an effective technique for active learning, and it has been used in a variety of applications, such as natural language processing, computer vision, and speech recognition.

Here are some of the benefits of using MQS:

- **Improved performance:** MQS can help to improve the performance of machine learning models by selecting the most informative data points for labeling.
- **Reduced need for labeled data:** MQS can reduce the need for labeled data, which can save time and money.
- **Increased interpretability:** MQS can help to increase the interpretability of machine learning models by providing insights into the decision-making process of the model.

Here are some of the challenges of using MQS:

- **Computational cost:** MQS can be computationally expensive, especially for large datasets.
- **Model bias:** MQS can be biased towards selecting data points that are similar to the data points in the training set.
- **Label noise:** MQS can be sensitive to label noise, which can lead to incorrect predictions.

Stream-based Sampling

Stream-based sampling is a type of active learning where the model selects the most informative data points from a stream of unlabeled data. The stream of unlabeled data is typically generated by a sensor or other device that is continuously collecting data. Stream-based active learning is particularly useful for applications where the data is constantly changing and it is not possible to store all of the data in memory.

How Stream-based Sampling Works

1. Initialize the labeled data set and the stream of unlabeled data.
2. Train a model on the labeled data set.
3. Select the most informative data points from the stream of unlabeled data.
4. Query the user for the labels of the selected data points.
5. Add the labeled data points to the labeled data set.
6. Repeat steps 2–5 until the desired performance is achieved.

Benefits of Stream-based Sampling

- **Reduced need for labeled data:** Stream-based sampling can reduce the amount of labeled data needed to train a model, which can save time and money.

- **Improved performance:** Stream-based sampling can lead to improved performance on a variety of tasks, such as classification, regression, and clustering.
- **Reduced risk of overfitting:** Stream-based sampling can help to reduce the risk of overfitting, which occurs when a model learns too much from the training data and is unable to generalize well to new data.

Strategies

Low Confidence Based Sampling

Low confidence based sampling is a technique used in active learning to identify the most informative data points for labeling. The idea behind this technique is that the model is more likely to make mistakes on data points that it is less confident about. Therefore, by selecting data points that the model is less confident about, we can improve the model's performance more effectively.

How Low Confidence Based Sampling Works

1. Train a model on the available labeled data.
2. For each unlabeled data point, compute the model's confidence in its prediction.
3. Select the unlabeled data point with the lowest confidence.
4. Query the user for the label of the selected data point.
5. Add the labeled data point to the training set.
6. Repeat steps 2-5 until the desired performance is achieved.

Benefits of Low Confidence Based Sampling

- Improved performance: Low confidence based sampling can help to improve the performance of machine learning models by selecting the most informative data points for labeling.
- Reduced need for labeled data: Low confidence based sampling can reduce the need for labeled data, which can save time and money.
- Reduced risk of overfitting: Low confidence based sampling can help to reduce the risk of overfitting, which occurs when a model learns too much from the training data and is unable to generalize well to new data.

Margin-Based Sampling

Margin-based sampling is a type of active learning where the model selects the most informative data points based on the margin of the model's predictions. The margin is the difference between the model's predicted probability of the correct class and the model's predicted probability of the most likely incorrect class. Data points with a small margin are more likely to be misclassified, and therefore are more informative for the model.

How Margin-Based Sampling Works

1. Initialize the labeled data set and the pool of unlabeled data.
2. Train a model on the labeled data set.
3. Compute the margin for each data point in the pool of unlabeled data.
4. Select the data point with the smallest margin.
5. Query the user for the label of the selected data point.
6. Add the labeled data point to the labeled data set.
7. Repeat steps 2-6 until the desired performance is achieved.

Benefits of Margin-Based Sampling

- **Reduced need for labeled data:** Margin-based sampling can reduce the amount of labeled data needed to train a model, which can save time and money.
- **Improved performance:** Margin-based sampling can lead to improved performance on a variety of tasks, such as classification, regression, and clustering.
- **Reduced risk of overfitting:** Margin-based sampling can help to reduce the risk of overfitting, which occurs when a model learns too much from the training data and is unable to generalize well to new data.

Challenges of Margin-Based Sampling

- **Computational cost:** Margin-based sampling can be computationally expensive, especially for large datasets.
- **Model bias:** Margin-based sampling can be biased towards selecting data points that are similar to the data points in the training set.
- **Label noise:** Margin-based sampling can be sensitive to label noise, which can lead to incorrect predictions.

Entropy-Based Sampling

Entropy-based sampling is a technique used in active learning to identify the most informative data points for labeling. It is based on the principle of information gain, which measures the amount of information that is gained by labeling a particular data point. The data point with the highest information gain is considered to be the most informative and is therefore selected for labeling.
Here's how entropy-based sampling works:

1. Train a model on the available labeled data.
2. For each unlabeled data point, calculate its entropy. Entropy is a measure of uncertainty, and it is higher for data points that the model is less certain about.
3. Select the unlabeled data point with the highest entropy.
4. Query the user for the label of the selected data point.
5. Add the labeled data point to the training set.
6. Repeat steps 2-5 until the desired performance is achieved.

The main benefit of entropy-based sampling is that it can help to improve the performance of a machine learning model by selecting the most informative data points for labeling. This can

lead to a reduction in the amount of labeled data that is required to train the model. However, there are also some challenges associated with entropy-based sampling. One challenge is that it can be computationally expensive to calculate the entropy of each unlabeled data point. Another challenge is that entropy-based sampling can be biased towards selecting data points that are similar to the data points in the training set.

Methodology

- Data Sets
    - Balanced Datasets
        - Iris
        - Digits
    - Imbalanced Datasets
        - BMI
- Used AL scenarios
    - Pool-Based Active Learning

Performance before active learning:

Iris:

Accuracy: 98%

F1 Score: 98%

Digits:

Accuracy: 93%

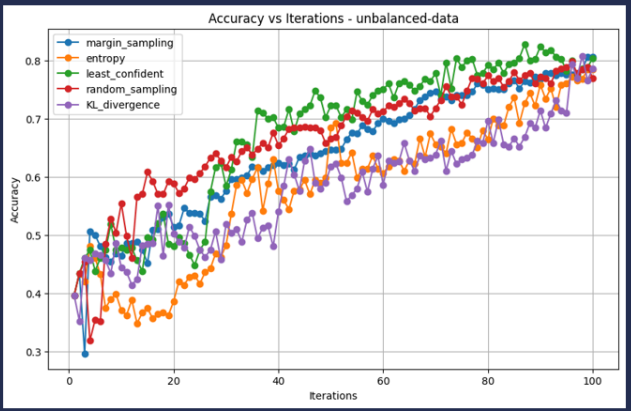
F1 Score: 93%

Enhancement Approach Performed for Unbalanced Dataset:

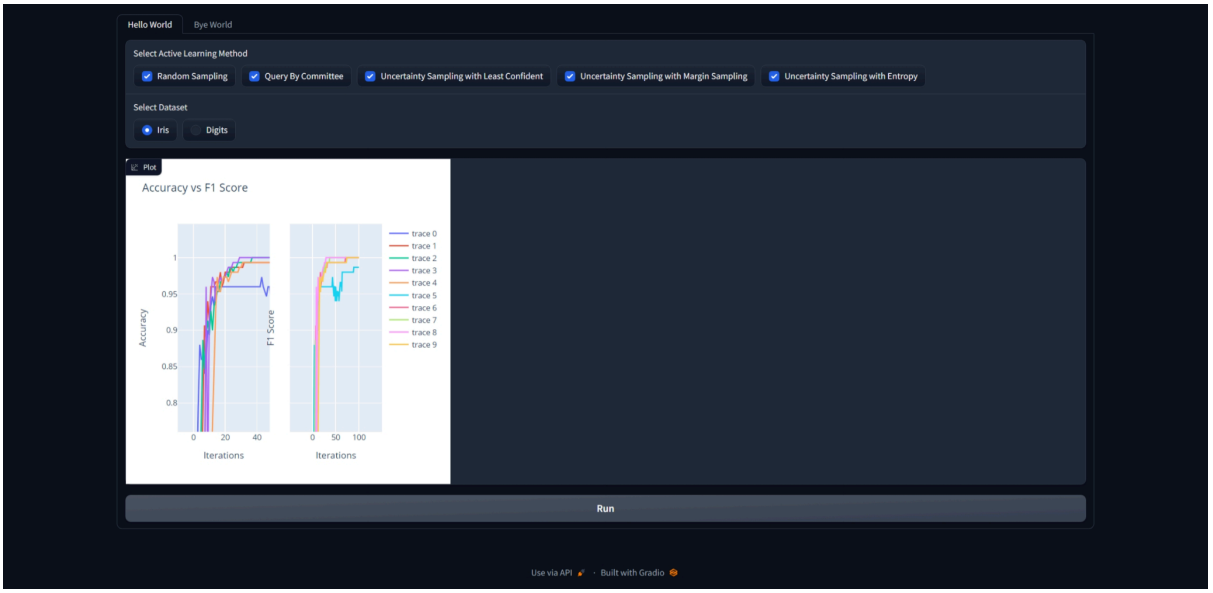
Applying the over-sampling technique has improved the performance of the model by a considerable percentage.

BMI unbalanced Dataset

Accuracies for a dataset: unbalanced-data
Method: margin_sampling, Accuracy: 0.8060
Method: entropy, Accuracy: 0.7840
Method: least_confident, Accuracy: 0.8040
Method: random_sampling, Accuracy: 0.7700
Method: KL_divergence, Accuracy: 0.7860



Deployment



Conclusion

Active learning represents a crucial strategy in the field of machine learning, particularly for deep learning models, by efficiently utilizing limited labeled data. The choice of sampling strategy can significantly impact the effectiveness of an active learning approach, making it essential to consider the specific needs of the task at hand.